From aligned models to trusted interfaces

Explainable health intervention and transparent health information seeking

Sun, X.

Publication date
2025

Document Version
Final published version

From Aligned Models To Trusted Interfaces:

Explainable Health Intervention And Transparent Health Information Seeking

Xin Sun

Xin Sun

From Aligned Models to Trusted Interfaces: Explainable Health Intervention and Transparent Health Information Seeking

Xin Sun

Funding
The work in this thesis was accomplished with support from the European Union Horizon 2020 program "TIMELY" project at the University of Amsterdam, the Netherlands (Grant ID: 101017424).

Cover by Xin Sun (with assistance from GPT-5 and Gemini 2.5)

ISBN: 978-94-6473-976-3

From Aligned Models to Trusted Interfaces: Explainable Health Intervention and Transparent Health Information Seeking

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 20 november 2025, te 10.00 uur

door Xin Sun
geboren te Jiangsu

Promotiecommissie

Promotor:	dr. J.A. Bosch	Universiteit van Amsterdam
Copromotores:	dr. J.M.S. de Wit	Tilburg University
	dr. A. El Ali	Centrum Wiskunde & Informatica
Overige leden:	prof. dr. B.J. Verschuere	Universiteit van Amsterdam
	prof. dr. T.B. Araujo	Universiteit van Amsterdam
	prof. dr. S. Ben Allouch	Universiteit van Amsterdam
	dr. T. Dingler	TU Delft
	dr. U.K. Gadiraju	TU Delft
	prof. dr. W.J. Kop	Tilburg University

Faculteit der Maatschappij- en Gedragswetenschappen

Contents

1
Introduction

The rapid advancements in large language models (LLMs) [1] have opened new frontiers in artificial intelligence (AI), revolutionizing fields such as natural language processing [2], conversational AI, and personalized interaction systems. These large language models, powered by billions of parameters, can generate human-like dialogue and adapt to diverse applications. In recent years, conversational agents, or chatbots, have received growing attention for their role in health interventions [3–5], offering scalable, on-demand support for mental health, chronic disease management, and health behavior change. Prior research has demonstrated that AI or LLM-driven conversational systems can provide personalized guidance, improve interactive engagement [6], and assist in behavior change [7], making them valuable tools in digital health interventions. Given the global shortage of healthcare professionals and the growing burden on mental health services, LLM-powered conversational agents present a promising solution to bridge accessibility gaps and extend healthcare support to underserved populations.

In both mental and physical health interventions, motivational interviewing (MI) [8, 9] is a well-established, evidence-based approach to health behavior change [10]. MI relies on core therapeutic skills such as reflective listening [11], eliciting intrinsic motivation, and empowering individuals to take ownership of their changes (i.e., fostering client autonomy). The ability to deliver MI through conversational agents offers promising potential to make behavior change interventions more accessible and scalable, particularly for individuals who face barriers to traditional care. The interactive and empathetic elements of MI align well with the nature of conversational agents, which are capable of building rapport through reflective

1

and motivational dialogue, for example, by paraphrasing a user's concerns to show understanding (reflective dialogue) or by encouraging self-efficacy and change talk (motivational dialogue). While early rule-based chatbots [12] have demonstrated value in structuring dialogue exchanges, they lack the flexibility to adapt to clients' diverse and evolving responses and rely heavily on expert-crafted content,because their responses are predetermined and must be manually designed to cover specific scenarios and context [13, 14].

With the advancement of LLMs, LLM-powered conversational agents on the other hand, dramatically improved at generating contextually rich and natural-sounding dialogue [15]. However, LLMs still lack inherent adherence to therapeutic principles. For instance, LLMs may offer advice prematurely or fail to elicit change talk, which contradicts key MI strategies that prioritize client autonomy and reflective listening. This can result in outputs that are inconsistent, ineffective, or even harmful [16] in sensitive contexts such as psychotherapy for health behavioral intervention. This highlights a critical need for approaches that align LLMs with domain expertise to ensure their outputs are not only contextually coherent and engaging but also therapeutically adherent, controllable, and explainable. Therefore, the work in this thesis investigates the approaches to align LLMs with domain expertise, specifically by integrating therapeutic strategies from evidence-based frameworks (i.e., motivational interviewing (MI) and cognitive behavioral therapy (CBT) [17]). Thus, alignment, in this work, refers to guiding or constraining the LLM's responses so that they follow established therapeutic strategies, rather than simply generating plausible-sounding dialogue. For example, instead of giving direct advice when a user expresses distress (which LLMs might do by default), an aligned model grounded in MI would respond with a reflective statement that encourages further self-exploration and supports empathetic engagement. MI's flexible, client-centered style emphasizes open-ended questions and reflective engagement [8, 18, 19], making it well-suited for evaluating LLMs' ability to support empathetic, context-aware dialogue. In comparison, CBT is a widely used, evidence-based approach that focuses on helping individuals identify and change unhelpful thought patterns and behaviors. It relies on structured, goal-oriented techniques, such as cognitive reframing, where a client is guided to re-interpret negative thoughts in a more balanced and constructive way [20]. This structured nature aligns well with LLMs operating under explicit therapeutic objectives. The work in this thesis investigates how aligning LLMs with these two complementary therapeutic frameworks can more readily achieve both therapeutic adherence and empathetically engaging interactions [21, 22].

In addition to LLM's potential for delivering MI-based psychotherapy for health behavior interventions, LLMs are increasingly poised to support another critical aspect of personal health: online health information-seeking [23, 24]. This form of digital engagement is often an integral part of the broader behavior change process, where individuals educate themselves about symptoms, treatment options, and lifestyle adjustments before or alongside health intervention. A significant portion of the population now utilizes digital resources to obtain medical advice, symptom assessments, and treatment recommendations [25, 26]. Studies have shown that individuals frequently use online search engines, health-related social media, and medical platforms to supplement traditional healthcare services. While these resources have democratized access to health information, they also introduce challenges related to varying content credibility, misinformation and difficulties in interpreting complex medical advice [27, 28]. The emergence of LLM-powered conversational search [29, 30] is transforming this landscape by providing interactive and tailored health information. Unlike traditional search engines that return a list of links, LLMs can provide coherent and synthesized responses accessed through conversational user interfaces (CUIs), mimicking human-like and context-aware dialogue interactions [31–33]. These LLMs can engage users through CUIs, offering immediate and contextually relevant information, which makes them particularly appealing in the domain of health information-seeking, where users often seek immediate, relevant, and easily digestible information and advice. As such, LLMs not only enable accessible health support but also play a growing role in shaping how people form health beliefs and make informed decisions, a key precursor to behavior change.

Despite these advantages, public confidence in AI-based health tools remains limited. A recent U.S. national poll [34] found that fewer than half of adults trust chatbots for health information, revealing a persistent trust gap in the general population. However, trust in traditional health authorities, such as government agencies and public health institutions, did not do much better depending on individual political views, reflecting societal polarization. In this context of widespread uncertainty and competing sources of online information, trust becomes a decisive factor in how individuals make health-related decisions. As prior work [35] has shown that trust significantly influences whether people accept or reject AI technologies, particularly in sensitive domains like healthcare [30, 36–38].

Across a century of scholarship, researchers have continued to propose various perspectives of what constitutes trust [39–41]. While no single definition fully captures all dimensions of trust, three core elements seem to stand out [39, 42]: (1) a positive expectation regarding the reliability or intentions of the trustee, (2) a willingness to accept vulnerability or risk, and (3) a context of uncertainty where the decision to rely on the trustee may impact one ' s outcomes. In the context of human-technology and human-machine interaction, Lee and See [43] further describe trust in automation as an attitude that the system will help achieve an individual ' s goals in situations involving uncertainty and risk. They argue that trust can extend beyond interpersonal expectations to include beliefs about a system's transparency, alignment with user goals, and competence (i.e., system's ability to perform its intended function accurately and reliably). For example, a health chatbot ' s competence involves delivering accurate and relevant health information, transparency means clearly communicating its limitations and information sources, and alignment with user goals refers to understanding and supporting the user ' s health-related needs or decisions.

With the rapid development of AI and LLMs, increasing research has focused on understanding user trust in AI-enabled systems [44], as well as identifying the progress, challenges, and future directions of AI technologies. For instance, Saleh et al. [45] highlight issues such as lack of transparency and difficulties in ensuring human oversight as key challenges for building trustworthy AI systems. Moreover, prior empirical work identifies multiple dimensions shaping trust in online information, including how information is presented (e.g., formatting and language style) [33], the agent's interaction style (e.g., tones, proactive asking and adaptability) [32, 33], the perceived credibility of the information source (e.g., whether the source is a known institution, expert, or anonymous user from social media) [46–48], and the transparency of content generation and labeling (e.g., whether it is disclosed that content was AI or human-generated and how that content was produced) [27, 49, 50]. While LLM-powered conversational search agents can foster search engagement and enhance the user experience [32, 33], they also raise concerns such as the credibility and reliability of information and potential over-reliance on LLM-generated content [51, 52]. Understanding how such factors influence trust is essential for ethical design of reliable and transparent LLM-powered health systems and user interfaces that users can confidently utilize for trustworthy health information-seeking.

Trust assessment has methodological dimensions. Whereas self-reported measures are widely used for their simplicity and directness, they are prone to biases such as social desirability and the Initial Elevation phenomenon, i.e., an inflated sense of trust at the beginning of an interaction [53–55]. These biases can compromise the reliability and validity of trust assessments. Behavioral and physiological measures may offer a complementary lens of implicit (i.e., automated or subconscious) responses for evaluating trust [56–58]. This measurement approach aligns with growing interest in Human-Computer Interaction, using physiological sensing to inform system design and evaluation [59]. For example, eye-tracking data, such as fixation duration and saccade patterns, can reflect cognitive engagement, while physiological signals like heart rate variability (HRV) [56, 60, 61] can indicate emotional arousal and cognitive burden. These multimodal signals may provide richer insights than self-report alone [61, 62].

To summarize the above, the integration of LLMs into the health context offers significant potential but also presents critical challenges, particularly in delivering the digital health behavioral interventions as well as trustworthy health communication for health information search and dissemination. Addressing these challenges requires methodological innovations in aligning LLMs with domain-specific expertise and guidelines to improve the therapeutic adherence, contextual appropriateness, and explainability. In addition, empirical research on the trust perception through both self-reports as well as behavioral and physiological sensing, as an innovative approach to trust assessment, is essential to inform the design of trustworthy, transparent, and human-centered LLM-powered healthcare systems and user interfaces for health behavioral intervention and health information seeking.

1.1 This thesis

Motivated by the above potentials and challenges, the work in this thesis sought to further study the potential of LLMs in delivering digital health behavioral intervention and health information. Also, it examined how human-perceived trust in LLM-powered health communication is shaped, particularly in the context of health information-seeking and dissemination. Additionally, this work explored implicit dimensions of human trust by incorporating behavioral and physiological sensing to capture implicit responses to LLM-generated health information. By aligning LLMs with domain expertise for explainability and understanding trust in LLM-powered health communication, this thesis introduces two key novelties: 1) developing the model-based approaches to enhance the explainability and controllability of LLMs for chatbot-delivered psychotherapy, ensuring alignment with domain expertise in sensitive contexts; and 2) empirically investigating how trust in LLM-generated health information is influenced by user interface, interaction types, and transparent UI cues. It also explores the use of physiological and behavioral assessment, in addition to self-report, to capture implicit dimensions of trust. By bridging LLM alignment with human-centered trust research, this thesis aims to contribute to the development of explainable, transparent, and trustworthy AI applications and user interfaces in LLM-powered health contexts for health intervention and information seeking. To that end, this thesis has two interrelated **research themes**:

(1) **Model**: Aligning LLMs with domain expertise to enable explainable and control-

lable chatbot-delivered psychotherapy for health behavior interventions.

(2) **Interface**: Understanding how trust perceptions are shaped when people interact with health information generated by LLMs and disseminated by LLM-powered conversational user interfaces with different modalities.

Research Theme 1: Model –Aligning LLMs with Domain Expertise for Health Intervention

This first theme investigates how LLMs can be aligned with domain-specific expertise to enable controllable and explainable chatbot-delivered psychotherapy for health behavior interventions. The work unfolds in three stages. First, a large-scale empirical study evaluated how well current LLMs (i.e., GPT-4, LLaMA, BLOOM) can generate MI-consistent reflective utterances. This study revealed both the strengths and limitations of LLMs in capturing the nuance of MI dialogue. Second, to enable more systematic exploration, a bilingual MI dataset annotated with Motivational Interviewing Skill Codes (MISC) [11, 63] was constructed with the MI conversations collected from real therapy sessions, providing a ground truth resource for subsequent training and evaluating LLMs in MI contexts. In parallel, a second dataset was developed consisting of expert-crafted therapeutic dialogues grounded in both MI and CBT. This dataset captures a wide range of behavior change scenarios and provides structured content for testing LLM alignment with expert-guided conversational flows and therapeutic strategies (i.e., MISC). Third, building upon these data resources, the research of this theme further proposed an expertise-driven alignment method (i.e., Script-Strategy Aligned Generation, SSAG) to enhance LLM adherence to therapeutic strategies [11, 64] and structures while maintaining conversational flexibility and engagement. Through empirical investigations, the research in this theme contributes practical methods and insights for developing LLM-powered psychotherapy chatbots that balance adherence to domain expertise with conversational flexibility and engagement.

RQ 1: *How well do LLMs perform in generating reflective utterances in motivational interviewing (MI) sessions? (Chapter 2)*

This question evaluates the extent to which LLMs can generate reflective dialogues by comparing LLM-generated reflections to human-authored ones, through human evaluation. It examines four assessment criteria: appropriateness, specificity, naturalness, and engagement, which are essential for fostering client self-exploration and motivation. This question establishes an initial understanding of LLM capabilities in generating MI reflections, highlighting both their potential and key limitations.

RQ 2a: Do expert-crafted dialogue scripts remain essential for chatbot-delivered psychotherapy in the era of LLMs? (Chapter 4)

RQ 2b: Can psychotherapy chatbots using script-strategy aligned generation (SSAG) achieve comparable conversational quality and therapeutic effectiveness to those using script-aligned generation (SAG)? (Chapter 3 & 4)

RQ 2c: To what extent can SSAG reduce the reliance on expert-scripted content for developing the psychotherapy chatbots? (Chapter 3 & 4)

These questions examine how LLMs can be aligned with domain expertise, i.e., expert-crafted dialogue scripts and MI strategies (i.e., MI skill codes [11]), to generate psychotherapeutic dialogues that are both therapeutically effective and conversationally engaging. Building on the constructed dataset resource (Chapter 3), we proposed and evaluated a script-strategy aligned generation (SSAG) approach (Chapter 4), where LLMs are first prompted to reason about the appropriate MI strategy before generating the corresponding dialogue response. Results from both automatic and comparative human studies show that this approach improves the therapeutic adherence and explainability of LLM-generated psychotherapeutic dialogues. This demonstrates that expertise alignment enhances the controllability and practical viability of LLM-powered psychotherapy for health behavior intervention, while highlighting the trade-offs between structure, flexibility, and engagement in LLM-powered psychotherapeutic health intervention.

Research Theme 2: Interface –Trust Perception in LLM-Powered Health Communication

The second theme in this thesis addresses another challenge by conducting a series of mixed-methods user studies. These studies examined how people perceive and evaluate trust in AI-generated health information, focusing on both the search agent (e.g., ChatGPT [65] vs. traditional search engine, Google) and the dissemination interface (i.e., text-based, speech-based, or embodied). Trust plays a pivotal role in the adoption of AI technologies in health, where perceived credibility, transparency, and ethical responsibility are essential. This theme also investigated whether users trust AI-generated content differently from human-generated information, and how disclosed source labeling (e.g., "AI-generated" vs. "human-generated") modulates these perceptions. Findings revealed that LLM-powered conversational agents generally foster higher trust than traditional search engines, but trust perceptions vary based on disseminated interfaces, source transparency, and disclosed labeling of the presented content. While LLM-generated health information was trusted more than human-generated content, human-labeled information was perceived as more credible than AI-labeled information, highlighting a complex trust paradox. To gain deeper insights, behavioral and physiological sensing, such as eye-tracking [66], electrocardiogram (ECG), and electrodermal activity (EDA) [62] signals, were used to capture the implicit trust-related responses. Ultimately, this theme provides insights into the design of user-centered, transparent LLM-powered applications by identifying how trust is shaped not only by content quality but also by interface design and transparent source labeling.

RQ 3a: How does people perceived trust in health information vary across different search agents (e.g., Google vs. ChatGPT) and dissemination interfaces (text-based, speech-based, or embodied)? (Chapter 5)

RQ 3b: How does inherent trust in these agents and interfaces influence the trust perception in health information? (Chapter 5)

RQ 3c: What factors contribute to trust perception across search agents and dissemination interfaces? (Chapter 5)

These subquestions focus on comparing trust perception across different conversational search agents and user interfaces, and analyzing how elements, such as interaction style, information presentation and dissemination interface with different modalities, affect users' trust in LLM-powered health information seeking.

RQ 4a: How do the actual source, disclosed label, and type of health information influence perceived trust? (Chapter 6)

RQ 4b: Can behavioral and physiological signals provide insights into trust perceptions toward human- and AI-generated health information? (Chapter 6)

These subquestions investigate whether users trust AI-generated health information differently from human-generated content, and how disclosed labeling (i.e., “AI-generated” vs. “Health Professional-authored”) influences users' trust perceptions. Behavioral and physiological sensing are additionally employed to capture how nuanced implicit responses could relate to user's trust perceptions in health information.

Together, these two themes of the thesis provide a cohesive exploration of LLMs in health-related domains, addressing both their alignment with domain expertise and their impact on user trust in LLM-powered systems or interfaces. Research in this thesis highlights the dual nature of LLMs: offering flexibility and scalability but requiring robust frameworks for alignment and thoughtful interface design to maximize their potential in digital health contexts.

1.2 Thesis Outline

This thesis comprises an introduction (Chapter 1), five research chapters (Chapters 2- 6), and a general discussion (Chapter 7). In Chapters 2-6, we explore two research themes: (1) aligning LLMs with domain expertise in delivering psychotherapy for health intervention, and (2) understanding human trust perceptions in health communication driven by LLM-powered conversational search. These chapters employ a combination of dataset creation, approach evaluation, and empirical experiments to address research questions.

The first research theme focuses on aligning LLMs with expertise-driven dialogue scripts and therapeutic strategies for motivational interviewing (MI) and cognitive behavior therapy (CBT)-based psychotherapy. **Chapter 2** evaluates the capability of LLMs to generate reflective utterances in MI. **Chapter 3** introduces the creation of an MI dataset annotated with MI skills codes [11], forming the basis for subsequent explorations. **Chapter 4** explores an expertise-driven approach for generating psychotherapeutic dialogues with controllability and explainability that are both contextually appropriate and therapeutically adherent.

The second research theme examines trust in health information-seeking facilitated by LLMs. **Chapter 5** investigates how people perceive trust in health information from traditional search engines compared to LLMs, disseminated through LLM-powered conversational user interfaces with different modalities. **Chapter 6** explores how the source of health information and the labeling of information as AI-generated or human-generated influence trust perceptions. Additionally, it incorporates behavioral and physiological sensing to uncover subconscious responses, providing deeper insights into how people perceive and evaluate LLM-generated health information.

Chapter 2 Before addressing the challenges of aligning LLMs with expertise for delivering MI, it is essential to initially evaluate their current capabilities in performing fundamental MI skills. This chapter examines how well LLMs can generate reflective utterances, an essential element of MI, by conducting human evaluations of LLM-generated reflections. The evaluation study involved layperson evaluators who assessed the quality of LLM-generated reflections in terms of four assessment criteria: appropriateness, specificity, naturalness, and engagement. The results highlighted both the strengths and limitations of LLMs in producing psychotherapeutically adherent and contextually relevant MI reflections, providing valuable insights into how LLMs can be utilized in MI applications.

Chapter 3 Building on the initial human evaluation of LLMs' ability to generate reflective utterances for MI in Chapter 2, this chapter focuses on the creation of a bilingual MI dataset to facilitate the deeper exploration of integrating LLMs with the domain expertise in MI. The dataset was annotated with MI skill codes [11], capturing key therapeutic behaviors in MI such as open-ended questions, affirmations, reflections, summaries, etc. This chapter detailed the process of dataset creation, including the collection of MI dialogues in Dutch from real clinics, the translation from Dutch to English, the annotation process, and the challenges encountered during the bilingual alignment. By providing a structured and reliable resource, the dataset can act as ground truth data and enable systematic exploration of LLMs' capacity to generate MI-adherent dialogues, laying the groundwork for subsequent exploration of how LLMs can be aligned with psychotherapeutic (MI) strategies for controllable and explainable generation.

Chapter 4 This chapter builds on the findings from Chapters 2 and 3 by exploring innovative approaches to align LLMs with domain expertise (i.e., expert-crafted dialogue scripts and MI strategies) for controllable and explainable psychotherapeutic dialogue generation in the context of MI and CBT. While LLMs have shown promise in producing dialogues, especially for generating the reflections in MI (Chapter 2), their inherent lack of controllability and explainability poses challenges, especially in the sensitive context of health behavioral intervention. To address this, MI strategies were employed as a framework for regulating the generation of a more controllable and explainable dialogue interaction. Specifically, this chapter introduced a two-step approach: first, instructing LLMs to predict the relevant MI strategies for the next conversational turn; and second, using these predicted strategies to guide the subsequent dialogue generation aligned to both the LLM-predicted MI strategies and the expert-crafted dialogue scripts. Through a series of automatic and human evaluations, this chapter examined the effectiveness of this alignment approach. The results highlighted the potential of strategy and script-aligned generation in making LLMs controllable and explainable for practical applications in the context of psychotherapy (i.e., MI and CBT) for health intervention.

Chapter 5 This chapter investigates how human trust in health information is shaped by traditional search engines compared to LLM-powered conversational agents, as well as by various conversational user interfaces with different modalities used to disseminate the LLM-generated health information. Through two mixed-methods studies, this chapter explored trust perceptions in the context of online health information seeking. Study 1 (N=21) compared trust in health information sourced from ChatGPT versus Google across different search tasks for health information, revealing significantly higher trust in ChatGPT, which demonstrates the potential of LLM-powered conversational search. Building on these findings, Study

2 (N=20) examined how different dissemination interfaces: text-based, speech-based, and embodied, powered by an identical LLM, affect trust in LLM-sourced health information, uncovering significant variations in trust across interface types. Interviews from both studies identified key factors influencing trust, including source credibility, interactive mode and information modality, search autonomy, and participants' prior knowledge. These findings highlighted the potential of LLM-powered conversational search for health information-seeking and provided actionable insights for designing LLM-powered health tools and interfaces in fostering trust.

Chapter 6 This chapter explores how information sources and labeling of sources influence trust in LLM-generated versus human-sourced health information, employing behavioral and physiological sensing to gain deeper insights into trust perception in online health information. With the proliferation of LLM-generated health content, often indistinguishable from human-sourced information, understanding how people trust such content has become critical. A mixed-methods approach was employed, including a survey study (N=142) and a within-subjects lab experiment (N=40), to investigate the effects of information source (Human vs. LLM), disclosed labeling (Human vs. AI) and health information type (General vs. Symptom vs. Treatment) on trust perception. The findings revealed that LLM-generated health information was trusted more than human-generated content, regardless of labeling, while the health information with human labels was trusted more than AI labels. Trust levels remained consistent across different types of health information. Behavioral and physiological sensing data, including eye-tracking and other indicators (i.e., EDA, ECG, and body temperature), were analyzed to predict people's self-reported perceived trust level and classify the information sources, achieving 73% accuracy in binary classification of trust level and 65% accuracy in source identification. These findings highlighted the role of transparency in information sources and labeling in modulating trust perception, demonstrating the potential of behavioral and physiological measures to verify trust levels and detect when additional transparency is needed.

Chapter 7 In this final chapter, we synthesize the findings from previous chapters to address the research questions and highlight the thesis's contributions as well as implications. We reflect on the challenges and opportunities of the research investigated in this thesis. Ultimately, we propose future research directions, such as expanding LLM alignment to broader contexts, longitudinal evaluation of AI-powered health systems in real-world scenarios, and refining the trust measurement methods specifically for human-AI interaction. This chapter envisions a future where LLMs are responsibly and transparently integrated into digital health interventions and communications.

1.3 Main Contributions

This thesis contributes to the fields of LLM-powered health in several ways, spanning data resources, methodological and empirical advancements. Methodologically, it introduces approaches for aligning LLMs with domain expertise for health behavioral interventions. Empirically, it provides evidence-based insights of human trust in LLM-powered health information seeking. Together, these contributions advance our understanding of how LLMs can be controllable and transparently integrated into LLM-powered healthcare systems, addressing both technical and human-centered challenges. By addressing these interconnected challenges, this thesis aims to bridge the gap between technical innovation and human-centered design in the era of LLMs for digital health intervention and communication. It envisions a future where LLMs are not only powerful tools for advancing health but also ethically aligned and trusted by the users they serve.

1.3.1 Dataset Contributions

(1) A bilingual dataset annotated with motivational interviewing (MI) skill codes (e.g., open-ended questions, affirmations, reflections, summaries) for training and evaluating LLMs in MI applications. (Chapter 3)

(2) A structured, expert-crafted dialogue dataset for health behavior intervention developed by health psychology experts, grounded in motivational interviewing (MI) and cognitive behavioral therapy (CBT). The dataset captures realistic therapeutic interactions and supports training and evaluation of LLMs in generating therapeutically coherent and script-aligned dialogues. (Chapter 4)

1.3.2 Methodological Contributions

(1) A novel approach for expertise-driven alignment (i.e., Script-Strategy Aligned Generation (SSAG)) of LLMs, instructing LLMs to predict MI strategies before generating the dialogues, enhancing the explainability and controllability of LLM generation in the context of MI and CBT for health behavior intervention. (Chapter 4)

1.3.3 Empirical Contributions

(1) A large-scale empirical human evaluation of LLMs' ability to generate reflective utterances in motivational interviewing (MI) based on the assessing criteria appropriateness, specificity, naturalness, and engagement. (Chapter 2)

(2) An empirical investigation into the effectiveness of the proposed expertise-driven alignment approach (i.e., Script-Strategy Aligned Generation, SSAG) for enhancing controllability and explainability of LLM in psychotherapy contexts. (Chapter 4)

(3) An empirical exploration of user trust in LLM-generated health information, comparing traditional search engines (i.e., Google) with LLM-powered conversational agents (i.e., ChatGPT), and investigating how conversational user interfaces (i.e., text, speech, embodied-based) influence user perceived trust in delivered LLM-generated health information. (Chapter 5)

(4) An empirical investigation on the impact of information source and transparency labeling (AI-generated vs. human-generated) on trust. Using behavioral and physiological sensing, the study demonstrates the potential of these implicit measures in predicting self-reported trust levels and accurately classifying information sources, offering valuable insights for objectively evaluating trust in AI systems. (Chapter 6)

1.4 Origins

The research chapters in this thesis are built upon the following publications.

Chapter 2 is based on the following paper:

- Xin Sun, Erkan Basar, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A.Bosch, Emiel Krahmer. How Well Can LLMs Reflect? A Human Evaluation of LLM-generated Reflections for Motivational Interviewing Dialogues. In Proceedings of the International Conference on Computational Linguistics (COLING 2025)

  XS and EB contributed to the study conception, formulated the research idea, implemented the codes, conducted the experiments, and did most of the writing; EK led the discussions and supervised the project; All authors offered valuable suggestions and contributed significantly to the text.

Chapter 3 is based on the following paper:

- Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos TP Dobber, Jos Bosch. Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)

XS and JhP contributed to the study conception, formulated the research idea and did most of the writing; XS implemented the codes and conducted the experiments; JTPD provided the raw data; All authors offered valuable suggestions and contributed significantly to the text.

Chapter 4 is based on the following two papers:

- Xin Sun, Jan de Wit, Zhuying Li, Jiahuan Pei, Abdallah El Ali, Jos A Bosch. Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy. (Journal) Proceedings of the ACM on Human-Computer Interaction (Computer-Supported Cooperative Work And Social Computing, CSCW); and
- Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, Jos A Bosch. Rethinking the Alignment of Psychotherapy Dialogue Generation with Motivational Interviewing Strategies. In Proceedings of the International Conference on Computational Linguistics (COLING 2025)

XS, JdW, ZyL and JAB contributed to the study conception, formulated the research idea; XS and XT implemented the codes, conducted the experiments; XS did most of the writing; JdW and JAB led the discussion and supervised the project; All authors offered valuable suggestions and contributed significantly to the text.

Chapter 5 is based on the following two papers:

- Xin Sun, Yunjie Liu, Jos A. Bosch, and Zhuying Li. Interface Matters: Exploring Trust Perception in Health Information from Large Language Models via Text, Speech, and Embodiment. (Journal) Proceedings of the ACM on Human-Computer Interaction (Computer-Supported Cooperative Work And Social Computing CSCW); and
- Xin Sun, Rongjun Ma, Xiaochang Zhao, Janne Lindqvist, Jan de Wit, Zhuying Li, Abdallah El Ali, Jos A.Bosch. From Agents to Interfaces: Understanding Trust in Health Information from Conversational Search. Under review by the Journal of Behaviour & Information Technology (BIT).

XS, RjM and ZyL contributed to the study conception, formulated the research idea; XS conducted the experiments and did most of the writing; XS, RjM, YjL and XcZ prepared the study materials and did the data analysis; JAB and AEA led the discussion and supervised the project; All authors offered valuable suggestions and contributed significantly to the text.

Chapter 6 is based on the following paper:

- Xin Sun, Rongjun Ma, Shu Wei, Pablo César Garcia, Jos A.Bosch, Abdallah El Ali. Understanding Trust Perception in AI-Generated Information with Behavioral and Physiological Sensing. Under review by the International Journal of Human-Computer Studies(IJHCS) with major revision.

  All authors conceived the study conception, formulated the research idea, offered valuable suggestions and contributed significantly to the text; XS prepared the study materials, conducted the data analysis, and did most of the writing; AEA led the discussion and supervised the project.

Part I

“Model”: Aligning Large Language Models (LLMs) with Domain Expertise for Psychotherapeutic Health Intervention

2

Evaluation of LLMs Generated-Reflections in Motivational Interviewing (MI)

This chapter is based on the following publication:

Authors: Xin Sun*, Erkan Basar*, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A.Bosch, Emiel Krahmer. (*equal contribution)

Original title: How Well Can LLMs Reflect? A Human Evaluation of LLM-generated Reflections for Motivational Interviewing Dialogues

Published in: International Conference on Computational Linguistics (COLING 2025)

Abstract

Motivational Interviewing (MI) is a counseling technique that promotes behavioral change through reflective responses to mirror or refine client statements. While advanced Large Language Models (LLMs) can generate engaging dialogues, challenges remain for applying them in a sensitive context such as MI. This work assessed the potential of LLMs to generate MI reflections via three LLMs: GPT-4, Llama-2, and BLOOM, and explored the effect of dialogue context size and integration of MI strategies in prompts for reflection generation by LLMs. We conducted evaluations using both automatic metrics and human judges on four criteria: appropriateness, relevance, engagement, and naturalness, to assess whether these LLMs can accurately generate the nuanced therapeutic communication required in MI. While we demonstrated LLMs' potential in generating MI reflections comparable to human therapists, content analysis showed that significant challenges remain. By identifying the strengths and limitations of LLMs in generating empathetic and contextually appropriate reflections in MI, this work contributed to the ongoing dialogue in enhancing LLM's role in therapeutic counseling.

2.1 Introduction

MOTIVATIONAL INTERVIEWING (MI) is an effective client-centered counseling technique designed to encourage behavioral change by helping clients explore and resolve ambivalence [67]. Reflective responses, which mirror or subtly rephrase clients' statements, are central to MI, deepening clients' motivation for behavioral change [67,68]. The empathetic reflections can enhance client engagement and therapeutic alliance, thereby influencing therapeutic outcomes.

Recently, there has been growing interest in how technology, particularly chatbots, can complement MI-based interventions [5, 69]. Besides their potential for scalability and cost-effectiveness, chatbots offer additional advantages including 24/7 availability, and the ability to provide anonymous and non-judgmental support. Traditional MI chatbots have relied on expert-written scripts and predefined rules to produce therapeutic dialogues [5, 69–72]. The reliance on scripted content restricts dialogue diversity and requires significant domain expertise and efforts on dialogue design. Several studies have attempted to improve this by generating MI reflections using templates [6, 73, 74]. However, these methods are limited by insufficient contextual understanding and an inability to replicate the depth of human empathy, which are crucial for effective MI.

Natural Language Generation (NLG) [75–77] with Large Language Models (LLMs) [78] marks a significant evolution from pre-scripted conversational MI applications, offering new possibilities for creating diverse, flexible, and MI-adherent dialogues by integrating MI expertise through in-context learning and few-shot capabilities [79]. However, employing NLG to automate MI reflections poses practical challenges. Therapeutic counseling requires that NLG technologies effectively handle the complex nuances of human communication, ensuring that reflections are not only contextually appropriate but also therapeutically accurate. The technical limitations of current LLMs, along with ethical considerations in automated therapeutic interactions, present significant obstacles [80–82]. Additionally, the effectiveness of LLM-generated reflections is heavily influenced by the prompts used. Therefore, there is a critical need for rigorous evaluation to assess how different prompts affect the generated MI dialogues by LLMs, ensuring they meet the high standards of contextuality and ethics required in MI. Given these challenges, we establish the following research questions:

(**RQ1**) Are LLMs capable of generating MI reflections with qualities comparable to human therapist reflections?

(**RQ2**) How does the size of the conversation context in prompts affect the quality of generated MI reflections?

(**RQ3**) Can the incorporation of MI strategies into LLM prompts enhance the quality of generated MI reflections?

We thereby conducted experiments using the proprietary model GPT-4 [83], the open-source model Llama-2 [84], and the open-science model BLOOM [85] to evaluate their effectiveness in generating reflections within the MI context. Utilizing the open-source MI dataset "AnnoMI" [86, 87] with human-human counseling dialogues, we assessed these LLMs' capabilities for generating MI reflections. After an automatic evaluation, we recruited 184 human evaluators to comprehensively assess the generated reflections based on the selected four criteria: appropriateness, specificity, naturalness, and engagement. By these evaluations, we investigated how well LLMs can reflect the nuanced communication required in MI settings.

This study evaluated the effectiveness of LLMs in generating MI reflections, with the goal of advancing empathetic, engaging, and effective conversational AI for psychotherapy. It compared the performance of leading LLMs across different prompting strategies to identify which combinations produce the most effective therapeutic communications. The extensive human evaluation was a core strength of this research, offering a robust measure of the practical effectiveness of AI-generated responses in psychotherapeutic settings. By highlighting the capabilities and limitations of various LLMs and prompting variants, this work provided valuable insights to both linguistic and psychological communities, laying a foundation for future advancements in LLM-enhanced MI.

2.2 Related Work

2.2.1 Generation of MI Reflections

Motivational Interviewing is a client-centered counseling technique that fosters health behavior change. Reflection is central to MI, where therapists mirror and empathize with clients' thoughts and feelings, crucial for building rapport and supporting effective MI therapy [88, 89].

Natural Language Generation [77] in the context of MI mainly involves the creation of reflective listening responses ("reflections"). The primary goal of these responses is to mimic the therapeutic efficacy of human therapists, who use reflections to strengthen rapport and elicit client motivation toward change [88, 89]. While MI also includes techniques such as open-ended questions, affirmations, and summaries, reflections are are particularly suited for NLG due to their structured yet context-sensitive nature. Other components, which are more interactive or context-dependent, pose greater challenges for generation and are addressed in later chapters in relation to their integration with reflections. Other MI components, though important, often involve more interactive or judgment-based decision-making that pose different challenges for automated generation. These aspects are discussed in later chapters of this thesis in relation to their integration with reflections.

NLG of reflections offers several benefits in MI. Firstly, it can provide consistent and immediate reflective feedback, which is crucial in MI. Additionally, there is limited availability of trained therapists in light of the high demand. NLG can handle a high volume of sessions simultaneously, increasing the accessibility of MI-based interventions. Despite the benefits, implementing NLG in MI has numerous challenges. The primary difficulty lies in the development of systems capable of generating genuinely context-aware and empathetic responses. Reflections must be tailored not just to the content, but also to the emotional subtext of the client. Moreover, ethical concerns arise regarding the appropriateness of responses, especially in sensitive scenarios [80, 82].

Prior work has been made in the field of NLG for generating MI reflections. Early work focused on rule-based approaches that utilized templates to mirror client utterances [73, 90]. Further previous studies employed machine learning approaches to produce more nuanced and contextual reflections [91–93], which rely on large datasets of therapist-client interactions to learn reflective techniques. Recent advancements in large language models, have opened new avenues for exploring the automated generation of MI reflections. These advanced models can rephrase what clients say, reflecting their words or even emotions, in ways that feel genuine and empathetic [94], showing promising results in enhancing client engagement.

2.2.2 Evaluation of NLG

The performance of NLG can be critically assessed through both automatic and human evaluations [95–97]. This process is essential for determining how effectively NLG systems can produce human-like, contextually appropriate, and engaging responses, which are crucial for the success of conversational agents in diverse applications.

Automatic evaluation metrics play a fundamental role in assessing NLG tasks. Metrics like BLEU [98], ROUGE [99], and METEOR [100] are commonly used to provide objective assessments of textual similarity between generated dialogues and references. More recently, embedding-based metrics like BERTScore [101] have been developed to capture semantic similarities more effectively than traditional metrics [95, 96].

In addition to the automatic assessment, human evaluation can provide vital insights into aspects that automated metrics might overlook, including fluency, coherence, relevance, and engagement [95]. Initial human evaluation methods often relied on simple Likert scales where evaluators rated conversations [95]. Recent advancements have introduced more sophisticated techniques like pairwise comparison and ranking-based approaches, such as Rank-based Magnitude Estimation (RankME [102]). Despite the benefits, human evaluation faces challenges such as high costs, time consumption, and variability based on subjective interpretations by evaluators. The absence of standardized protocols complicates comparisons across different studies. Nonetheless, human evaluation remains indispensable for understanding how dialogue generation can emulate human conversational nuances.

2.3 MI Reflection Generation

2.3.1 Conversation Contexts

We used a publicly available MI dataset, "AnnoMI" [86, 87], which was compiled by transcribing the English spoken dialogues between therapists and clients on various topics such as alcohol and nicotine consumption. The data were annotated based on Motivational Interviewing Skills Code (MISC), which is a coding scheme providing a systematic way for assessing MI-adherent behaviors in therapist-client interactions [103, 104], such as the therapist's behaviors (e.g., reflection, question) and the client's behaviors (e.g., change talk, sustain talk). By using the MI-adherent dialogues of AnnoMI dataset, we created conversation contexts that consist of up to 5 dialogue turns between a therapist and a client where the final therapist response is labelled as a "reflection" behavior, as shown in Table 2.1.

To prepare the data for human evaluation, we first filtered contexts to reduce redundancy and ensure diversity among utterances, balancing representativeness with feasibility for manual analysis. This filtering was based on automatic evaluation results (see Section 2.4.1) and manual content analysis.[1] From the resulting 194 contexts, we randomly selected 160 for inclusion in the human evaluation study.

Utterances	MISC
Client: I guess it's because I know that I need to do it to lose weight	CT
Therapist: So, you realize, again, that if you decrease the amount of juice you're taking in, you're gonna decrease your weight you're gonna feel better	RF

Table 2.1: An example of an exchange between client and therapist in the AnnoMI dataset. "RF" stands for reflection and "CT" stands for change talk.

[1]The sampling process is detailed in Appendix 2.A.

2.3.2 Large Language Models

We employed three prominent LLMs in our study[2]. **GPT-4** [83] is widely accepted as the state-of-the-art LLM that is a proprietary close-source model. Therapeutic counseling, however, often deals with sensitive and personal information, making it important to consider using open-source models that can be operated on internal hardware. Thus, we also incorporated **Llama-2** [84] to our experiments, as it is a well-known open-source model developed as a competitor of GPT-4. Moreover, [105] showed that the extent to which the LLMs are open in practice fluctuates substantially, from the lack of scientific documentation to transparency in data collection. Therefore, we also experimented with **BLOOM** [85] as it remains one of the most open LLMs[3] and was developed by following open-science principles.

2.3.3 Prompting Strategies

We first utilized the following base prompt as the "task instruction" to guide the LLMs to generate the MI reflection inspired by prior work [106,107]:

> As a therapist of Motivational Interviewing, please generate the next appropriate utterance based on the dialogue history. Restriction: you MUST NEVER ask new questions.

Subsequently, we aimed to explore the effects of 1) the conversation context size and 2) the inclusion of MI strategies on the quality of generated reflections. We created four different prompting strategies from the combinations of the following prompting features[4]:

1-turn: the preceding 1 turn of dialogue is given as the conversation context.

5-turns: the preceding 5 turns of dialogue is given as the conversation context.

Full-MI: the MI strategies are incorporated as additional instructions. Specifically, each utterance in the conversation is assigned a MISC code, with corresponding definitions and examples provided. The LLMs are instructed to generate the next utterance according to the specified MISC code of "Reflection".

Partial-MI: the MI strategies are *not* incorporated within the prompt.

2.4 Evaluation Approaches

2.4.1 Automatic Evaluation

To objectively evaluate the effects of different prompting strategies on LLM-generated MI reflections, we utilized well-established automatic evaluation metrics:

---

[2] Implementation details are given in Appendix 2.B.

[3] According to the Opening Up ChatGPT list on 31 May 2024 https://opening-up-chatgpt.github.io/.

[4] The prompt template is provided in the Appendix 2.F.

text length, BERTScore, and BLEURT. The average length of generated text indicates verbosity or conciseness, crucial in MI sessions where clarity and brevity are key. BERTScore [108] uses BERT's contextual embeddings to evaluate the semantic similarity between texts, providing a more nuanced assessment than ROUGE [99], which relies solely on text overlap. BLEURT [109] combines traditional metrics with BERT's embeddings and is trained on human ratings, making it well-suited for evaluating the subtleties in LLM-generated MI reflections. We calculated the metrics (i.e., BERTScore and BLEURT) between each pair of generations produced from all six combinations of four prompting strategies for each of the three LLMs (e.g. GPT-4 1-turn Partial-MI vs GPT-4 5-turns Partial-MI). To focus human evaluation on diverse outputs, we excluded conversation contexts[5] where the generations were highly similar based on the average BLEURT and BERTScore.

Model	1-turn vs. 5-turns			Full-MI vs. Partial-MI		
Prompt	BERTScore	BLEURT	Lengths	BERTScore	BLEURT	Lengths
BLOOM	0.89	0.49	15 vs. 15	0.88	0.48	15 vs. 15
Llama-2	0.89	0.53	24 vs. 22	0.89	0.50	20 vs. 26
GPT-4	0.89	0.54	30 vs. 30	0.88	0.48	22 vs. 39

Table 2.2: The average BERTScore, BLEURT score, and text lengths for the MI reflections generated by the three selected LLMs. Comparisons are made between two prompting strategies per model. The average length of corresponding human reflections is 22 words.

2.4.2 Human Evaluation

Experimental Design

We recruited 184 participants through the Prolific crowd-sourcing platform, requiring fluency in English and being over 18 years old. These participants, residing in 25 different countries, were equally divided between men and women, with an average age of 31. Each participant evaluated reflections for both independent and ranking evaluations across 3 randomly assigned conversation contexts. Eventually, each context was evaluated by at least 3 different participants. We presented conversation contexts with 5 turns to the participants. [110] demonstrated that non-experts can evaluate MI reflections as effectively as MI experts. Following their findings, we recruited non-experts for our study to examine their perception of the generated reflections. We employed a Balanced Latin Square counterbalance measure to systematically rearrange the model positions at each context, to prevent potential order effects that could arise from presenting the models in fixed sequences [95].

[5] The sampling process is detailed in Appendix 2.A.

Independent Evaluation

The first part of the human evaluation focused on independently evaluating the quality of LLM-generated and human reflections, based on the provided conversation contexts. Participants were given a single reflection at a time and asked to evaluate it based on the following four distinct criteria at once. **Appropriateness** measures whether the reflection would be (emotionally and morally) appropriate if it is actually uttered to a client after the given conversation. **Specificity** is to understand whether the reflection contains elements from the client ' s previous response. **Naturalness** assesses whether the reflection sounds like it could have been uttered by a person. **Engagement** to see whether the reflection could provide the opportunity for further conversation and could increase the engagement.

The criteria were chosen by considering their relevance and importance to therapeutic counseling and their common usage in the NLG field[6]. For instance, we look into *appropriateness* because inappropriate reflections can hinder the clients' progress towards their behavior change goals [67]. Similarly, clients' *engagement* during counseling shown to be closely linked to their therapeutic progress [112], and striking a balance between *specificity* and genericness in reflections is crucial to keep a conversation interesting [113]. Likewise, ensuring *natural-sounding* reflections is as essential in order to maintain engagement and encourage ongoing interactions during counseling.

At the start of the survey, the participants were given a brief description along with mock-up examples of both positive and negative responses for each criterion[7]. A 7-point Likert scale gradually ranging from Strongly Disagree (-3) to Strongly Agree (3) was implemented [114].

Ranking Evaluation

The second part of the human evaluation aimed to compare the overall quality of generated and human reflections by directly ranking them. We utilized the RankME approach [102], which eliminated the need for multiple pairwise comparisons by having evaluators indicate the contrast between a pre-selected reference text and all target texts simultaneously through a process of magnitude estimation. In our study, we selected human reflections as the reference text (scored as 100) and asked our participants to assign a score to generated reflections given the human reflection and the conversation context.

We utilized TrueSkill [115] to ascertain the overall ranking among the models with their various prompting strategies. TrueSkill computed a mean rating value as the def-

[6] The criteria were proposed and elaborated in a previous publication [111].
[7] The survey is provided in the supplementary materials.

inite score for each combination by comparing them pairwise, where the higher-rated ones were considered the winners over the lower-rated ones. Following the original work, we set the initial rating to 25.

2.5 Results

2.5.1 Automatic Evaluation

Table 2.2 illustrates the results from the automatic evaluation. The automatic evaluation of three LLMs across prompting strategies ("1-turn vs. 5-turn" and "Full-MI vs. Partial-MI") employed BERTScore and BLEURT metrics alongside the average length of generated utterances. All LLMs consistently achieved an average BERTScore of around 0.89, indicating high semantic similarity in all setups. BLEURT scores were slightly varied, with Llama-2 and GPT-4 showing higher similarity in the generation than BLOOM. In terms of generation length, GPT-4 generated notably longer texts in the partial-MI setup compared to others, suggesting differences in handling extended MI contexts. These results suggested modest differences among the LLMs in automatic evaluation metrics.

2.5.2 Independent Human Evaluation

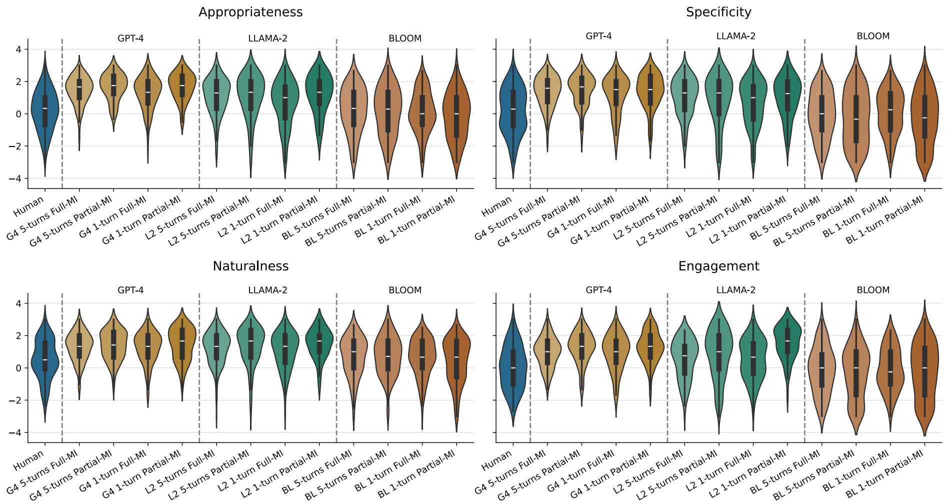


Figure 2.1: Violin plots display the distribution of 7-point human evaluation scores for each model and prompting strategy combination across each criterion, highlighting key statistics such as the median (white dashes) and the interquartile range (thick black bars), while also visualizing the score density of the variables, with wider sections representing higher density. Note that while our actual data falls within the range of $(-3, 3)$, the density estimations in the violin plots extend to $(-4, 4)$ due to the calculation of a continuous probability.

Figure 2.1 visualizes the distribution of 7-point evaluation scores, ranging from

-3 to 3, for each model and prompting strategy across each criterion, where a wider range on the graph indicates a larger score density. We observed that GPT-4 reflections received positive scores more frequently than negative ones, as evidenced by the short and narrow tails in Figure 2.1. Meanwhile, the human reflections had more balanced distributions across the criteria with the wider ranges being closer to zero compared to GPT-4. Moreover, Llama-2 garnered similar score distributions as GPT-4, except for the engagement criterion where Llama-2 1-turn partial-MI gathers higher positive scores more frequently. Finally, Figure 2.1 also indicates that BLOOM reflections received scores that were distributed similarly to human reflections but with more frequent negative scores, especially in specificity and engagement criteria, as shown by the wider tails.

A one-way ANOVA revealed the significance of the effect for all four criteria (appropriateness: $F(12, 147) = 46.27, p < .001$; specificity: $F(12, 147) = 38.49, p < .001$; naturalness: $F(12, 147) = 20.51, p < .001$; engagement: $F(12, 147) = 41.04$, $p < .001$). Tukey's HSD post-hoc test for multiple comparisons indicated the ratings given to all variations of GPT-4 reflections are significantly higher ($p < .05$) than human reflections across all criteria[8]. Likewise, the variations of Llama-2 reflections were significantly rated higher ($p < .05$) than the human reflections across all criteria, except that the 1-turn full-MI variation showed insignificant results in appropriateness and specificity. The difference between all variations of BLOOM reflections and human reflections was insignificant across all criteria. The results so far provided insights to answer **RQ1**.

Dimension	5-turns	1-turn	Full-MI	Partial-MI
Appropriateness	$\mu = 0.94, \sigma = 1.31$	$\mu = 0.79, \sigma = 1.37$	$\mu = 0.79, \sigma = 1.28$	$\mu = 0.93, \sigma = 1.39$
Specificity	$\mu = 0.73, \sigma = 1.46$	$\mu = 0.66, \sigma = 0.43$	$\mu = 0.70, \sigma = 1.35$	$\mu = 0.69, \sigma = 1.54$
Naturalness	$\mu = 1.11, \sigma = 1.11$	$\mu = 1.04, \sigma = 1.16$	$\mu = 1.02, \sigma = 1.09$	$\mu = 1.13, \sigma = 1.17$
Engagement	$\mu = 0.51, \sigma = 1.42$	$\mu = 0.66, \sigma = 1.40$	$\mu = 0.45, \sigma = 1.29$	$\mu = 0.72, \sigma = 1.51$

Table 2.3: Means (μ) and standard deviations (σ) of each prompting feature calculated per criterion based on the ratings provided by the evaluators for all models.

We performed multiple paired samples t-tests across the 4 criteria to compare 1) including 1-turn vs 5-turns in the prompt to answer **RQ2**, and 2) utilizing full-MI vs partial-MI instructions to answer **RQ3**. In Table 2.3, we see that 5-turns reflections were rated significantly more appropriate than 1-turn reflections ($t(11) = 2.363, p = .018$), and partial-MI reflections were more appropriate than full-MI reflections ($t(11) = 2.263, p = .024$). For the specificity criteria, there was no significant difference between 5-turns and 1-turn ($t(11) = 0.965, p = .335$),

[8]Visualised in Appendix 2.C, Figure 2.C.1.

2

or partial-MI and full-MI ($t(11) = -0.146, p = .884$). Partial-MI reflections were rated as more natural than full-MI reflections ($t(11) = 2.204, p = .028$) but the difference in naturalness between 5-turns and 1-turn reflections was insignificant ($t(11) = 1.191, p = .234$). Finally, 5-turns reflections were rated as less engaging than 1-turn reflections ($t(11) = -2.313, p = .021$), and partial-MI reflections were found significantly more engaging than full-MI reflections ($t(11) = 4.205, p < .001$).

2.5.3 Ranking Human Evaluation

We utilized TrueSkill to calculate a mean rating value (μ) and standard deviation (σ) for each model and prompting strategy based on the rankings given by the human evaluators. Figure 2.2 shows that only GPT-4 with 5-turns partial-MI ($\mu = 29.60, \sigma = 0.90$) reflections were ranked significantly higher than the human ($\mu = 26.98, \sigma = 0.85$) reflections, and GPT-4 with 1-turn partial-MI ($\mu = 26.72, \sigma = 0.85$) reflections showed no significant difference with human reflections. Reflections generated by other models and prompting strategies were ranked significantly lower than the human reflections.

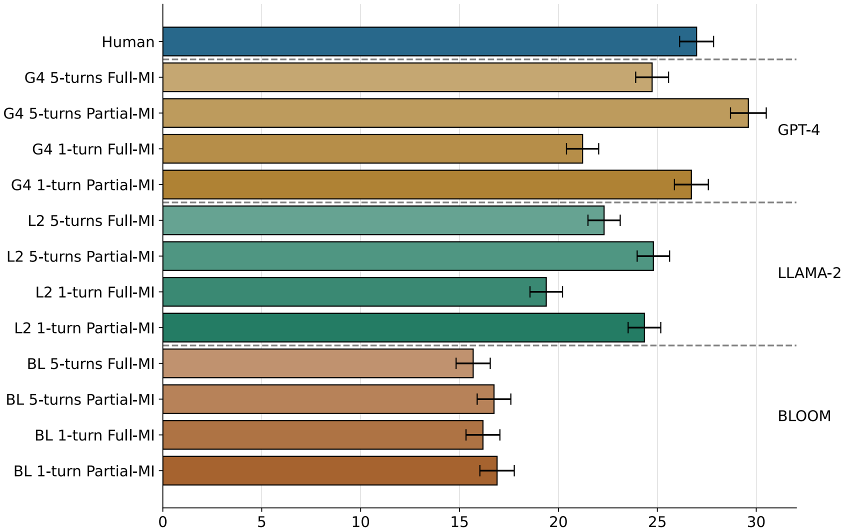


Figure 2.2: TrueSkill mean rating values (μ) computed for each model and prompting strategy based on the evaluators' rankings. The error bars indicate the standard deviation (σ).

When comparing the prompting strategies within the GPT-4 model, we observed that 5-turns partial-MI ranks higher than 1-turn partial-MI, followed by 5-turns full-MI, and then 1-turn full-MI ($\mu = 19.38, \sigma = 0.82$), all significantly. Moreover, an identical outcome applied to the prompting strategies for Llama-2, except that there was no significant difference observed between the 5-turns partial-MI ($\mu = 24.80, \sigma =$

Client: Well, I know my drinking has increased a little since I got divorced and moved to [location], but I don't think it's that much of a problem.
Therapist: It doesn't feel like a problem to you right now. **GPT-4 1-turn Partial-MI**: It seems like you've gone through some significant life changes recently with your divorce and move to [location], which might have impacted your drinking habits. acknowledging this is an essential first step. just remember, it's not about labeling yourself as a problem drinker, it's more about understanding the relationship you have with alcohol and how it might be affecting your life. **Llama-2 1-turn Partial-MI**: Well, let's take a closer look at this drinker's pyramid. It's true that you're not at the very top of the pyramid, but you're still in the harmful range. And it's interesting that you mention your recent divorce and move to [location]. Can you tell me a little more about how your drinking has changed since those events? **Llama-2 1-turn Full-MI**: It sounds like you're downplaying the significance of your drinking habits.

Table 2.4: An example context showing the client's final utterance and select few reflections rated variously by the evaluators. Full conversation context is provided in Appendix 2.G, Table 2.G.1.

0.82) and the 1-turn full-MI ($\mu = 19.38, \sigma = 0.82$). Changing the prompting strategy displayed no significant difference for the ranking of the reflections generated by BLOOM.

We observed that GPT-4 consistently ranks significantly higher than Llama-2 across all prompting strategies utilized. For instance, GPT-4 ($\mu = 24.73, \sigma = 0.83$) was ranked higher than Llama-2 ($\mu = 22.30, \sigma = 0.81$) when both were given 5-turns full-MI in their prompts, as well as with all other prompting strategies. Each BLOOM reflection was ranked significantly lower than the rest of the models.

A Kruskal-Wallis test confirmed the statistical significance of the variations in rankings among the reflection types ($H(12) = 624.304, p < .001$).

2.6 Discussion

In this study, we sought to evaluate the quality of MI reflections generated by three LLMs: GPT-4, Llama-2, and BLOOM, in comparison to human reflections. We explored the effects of utilizing different prompting features; shorter vs longer conversation contexts and succinct vs detailed MI instructions. We conducted two separate human evaluations, independent and ranking, and came to conclusions by analyzing both outcomes.

The independent evaluation results demonstrated that prompting LLMs with longer conversation contexts leads to generating more appropriate but less engaging reflections. The ranking evaluation results also displayed a preference over the reflections generated on longer contexts. This outcome suggested that LLMs benefit from the additional information provided by longer contexts, allowing them to generate more precisely worded outputs, which answer **RQ2**. For more engagement-focused applications, shorter contexts, which focus on capturing the essence of the conversation without overloading the LLMs, may be preferable.

The results of our independent evaluation study showed that as more detailed MI instructions are included in the prompts provided to the LLMs, appropriateness, naturalness and engagement evaluation scores for the generated reflections significantly decline. Likewise, the ranking evaluation results showed that prompts with detailed MI instructions can have a negative impact on the perceived quality of the generated reflections. This phenomenon can be attributed to the models' tendency to produce reflections that follow strict standards and miss the affective tone when provided with excess instructions, thus answering our **RQ3**. Instead of providing specific instructions, it may be more beneficial to allow the LLMs to interpret open-ended prompts. This aligned with the core spirit of motivational interviewing, which relies on empathetic listening and client-centeredness.

The independent evaluation results indicated that LLMs can generate reflections that are perceived as appropriate, specific, natural, and engaging at comparable levels to human reflections. Overall the ranking evaluation results displayed a preference for human reflections over most of LLM-generated ones. This could imply that evaluators examine the reflections for qualities beyond our four criteria during the ranking evaluation. Moreover, both independent and ranking evaluations suggested that GPT-4 (5-turns Partial-MI) reflections seemingly outperform the human reflections, answering **RQ1**. However, these results illustrated the expectations and perceptions of the non-expert evaluators, without examining the professional standards of MI. Moreover, AnnoMI dataset consists of in-person counseling during which body language, facial expressions, and gazing may be part of the communication more than the uttered reflections. Human therapists can potentially tailor their reflections towards their clients, resulting in high variation in their styles and perceptions, which may explain the diversity of human reflection scores in independent evaluation.

Concerning these results, content analysis revealed that more formulaic and elaborate reflections may be judged as more appropriate. In the context in Table 2.4, the client admits their drinking increased due to recent stressful events, but it is not yet at a level that could cause serious health problems. The human reflection on this was found not appropriate ($\mu = -2.6$) by all three evaluators. For the same context, GPT-4 (1-turn Partial-MI) generated a more complex and elaborate reflection which was rated with the highest level of appropriateness ($\mu = 3$). Llama-2 (1-turn Partial-MI) generated a reflection similar to GPT-4 reflection in structure and style, and was evaluated as appropriate ($\mu = 3$). However, LLMs also contain the risk of generating more confrontational reflections despite the instructions to follow MI approaches. For example, for the same context, Llama-2 (1-turn Full-MI) generated a reflection missing the empathetic tone and sensitivity required in MI.

Client: I mean, I don't think I would be, you know, drinking as much or going to parties as much if I'm studying for school, or in college, or trying to get in law school and things like that.	
Therapist: Okay, so, the alcohol is something that might make it harder to reach your goals?	
G4 5-turns Partial-MI: I really admire your commitment towards your future goals and your understanding about how certain behaviors could distract from those goals. you seem to have a really clear vision for your future which is truly inspiring.	
G4 1-turn Full-MI: So, in your view, pursuing education and focusing on your career would naturally steer you away from heavy drinking and partying. it's like you see these goals as your pathway to change.	

Table 2.5: An example context showing potential subtle nuances lacking in generated reflections. Full conversation context is provided in Appendix 2.G, Table 2.G.2.

Further content analysis showed that generated reflections may lack subtle nuances found in human reflections. For instance, in the context in Table 2.5, the client indicated that focusing towards a future goal may reduce drinking. The human reflection urged the client to consider the impact of drinking on reaching future goals. Although GPT-4 (5-turns Partial-MI) reflection appeared to aim for the same outcome as human reflection, it assumed the client already acknowledges that drinking prevents reaching future goals and praises this sentiment, thus hindering further self-reflection. For the same context, GPT-4 (1-turn Full-MI) generated a reflection that aligns with client's statement but overlooked the chance of self-reflection on the current drinking habits.

These findings highlighted that LLMs are capable of generating reflections that fulfill the expectations of non-expert human judges. The utilization of LLMs could benefit various applications, such as enriching MI reflection sets for hybrid response generation in chatbots [116]. However, they were not substitutes for trained therapists in MI or other sensitive areas, and should be used with caution, particularly when emotional safety and nuanced understanding are crucial.

2.7 Limitations

The chosen criteria (appropriateness, specificity, naturalness, and engagement) may not capture all necessary dimensions of effective reflections in MI. For instance, the empathy level and therapeutic impact of the reflections could also be important evaluation factors, which should be examined in further research. Despite efforts to standardize human evaluations by providing examples and definitions, human evaluators may have differing interpretations of these criteria, leading to inconsistencies in scoring. Moreover, we recruited many individuals from various countries, who may not be native English speakers, which could have influenced our evaluation. Likewise, the human evaluations were conducted on a sampled subset of AnnoMI data, which may have influenced our results.

Our study was only focused on generating reflections for provided scenarios. Whether LLMs can conduct complete therapy sessions is not investigated within the scope of this study. While we acknowledge the potential advantages of employing chatbots in therapy, we only view this application as feasible in certain circumstances, such as acting as a support tool or serving as a training resource. When inspecting the results of this study, the readers should refrain from assuming that the LLMs possess the capability to substitute human therapists or conduct virtual therapy sessions autonomously.

2.8 Conclusion

We evaluated the capability of three LLMs to generate reflective responses in MI and examined how conversation context size and inclusion of detailed MI instructions in prompts affect their performance. A series of human evaluations showed that LLMs produce reflections with qualities comparable to those of human therapists. Content analysis further revealed that the LLMs contain the risk of generating reflections that lack emotional depth and nuance required for MI conversations. Additionally, we found that the size of the conversation context and adding detailed MI instructions to prompts impact different evaluation criteria in various ways. This study offers a comprehensive evaluation for MI reflections and highlights the challenges and opportunities of using LLMs in sensitive domains like therapeutic counseling. Future research should involve MI experts as evaluators, incorporate additional metrics like empathy and therapeutic alliance, and explore other strategies for embedding MI principles into LLMs to expand our understanding of the capabilities of LLMs in MI contexts.

Ethical Review

Before conducting our experiment, our institution's ethics board reviewed and identified our study being in accordance with ethical standards[9]. The individuals who participated in our study have been provided with prior information regarding the task, research objectives, workload, compensation, our privacy protocols, and our intended utilization of the collected data in research. If participants did not provide consent, they were automatically restricted from reaching the survey. No personally identifiable information was retained after the study concluded. The participants were compensated with £7 per hour.

[9] Established by the Ethics Committee of Social Sciences at Radboud University and registered with the reference number ECSW-LT-2023-9-12-68541.

Acknowledgements

This project is partly financed by the Dutch Research Council (NWO) with project number 406.DI.19.054 and European Commission in the Horizon H2020 scheme with agreement ID 101017424.

Chapter appendix

2.A Conversation Context Sampling

Due to the significant costs usually involved, utilizing the complete set of conversation contexts during human evaluation was not possible. Instead of a randomized selection on the full set, we chose to implement an informed sampling process in an attempt to increase the efficiency of the human evaluations. First, to focus human evaluation on diverse outputs, we exclude conversation contexts where the generations are highly similar. Specifically, we filter out contexts where, for at least two LLMs, more than three out of six generation pairs have similarity scores higher than the average BERTScore (0.88) and BLEURT score (0.48), indicated in Table 2.2 in Section 2.4.1. This filtering ensures that human evaluators assess generations that are sufficiently different. Further, contexts where discarded if at least 3 of their generated reflections were shorter than 4 words or longer than 80, to make room for contexts with more meaningful content in their generated reflections. Finally, we have manually filtered the conversation contexts based on the content of the 5 turns conversation contexts, such as human therapist reflection, following the set of rules below:

- The final therapist reflection is too short, too vague, or a small confirmation (e.g. “I understand”).
- The final therapist reflection is bisected, where the rest is shifted to the previous or next conversation context. Because the original dialogues were in-person speeches, the therapist utterances may be halved when the client backchannels while the therapist speaks.
- The client is listening and backchanneling more than contributing to the conversation while the therapist summarizes the session.
- The final therapist reflection focuses on information given by the client outside of the 5 turns we are utilizing. Hence, the conversation contexts given to the LLMs do not contain this information. This often occurs when the therapist is starting to summarize the session in the form of a reflection.
- More than one generated reflections meaninglessly repeat a therapist utterance from the conversation context and are longer than three words.

2.B Implementation Details of Generations with LLMs

We utilized the June 2023 edition of GPT-4, coded as `gpt-4-0613`, chat version of Llama-2 with 70B parameters, coded as `Llama-2-70B-chat-hf`, and 176B parameter version of BLOOM, namely `bloom-176b`. We used `openai` Python library to generate with GPT-4, and `requests` library to send requests to the HuggingFace API to generate with Llama-2 and BLOOM models. We opted for default hyperparameters, including the temperature as default 1 to control the randomness of generation. The models were used in compliance with their respective licenses and terms at the time of the study. OpenAI provides a Terms of Use. Llama-2 is licensed by META. And BLOOM is authorized under BigScience RAIL License v1.0.

---

[9] `https://platform.openai.com/docs/models/gpt-4`
[9] `https://huggingface.co/meta-llama/Llama-2-70b-chat-hf`
[9] `https://huggingface.co/bigscience/BLOOM`
[9] `https://api-inference.huggingface.co`
[9] `https://openai.com/policies/terms-of-use`
[9] `https://ai.meta.com/llama/license/`
[9] `https://huggingface.co/spaces/bigscience/license`

2.C Tukey's HSD Post-hoc Test Results Visualized

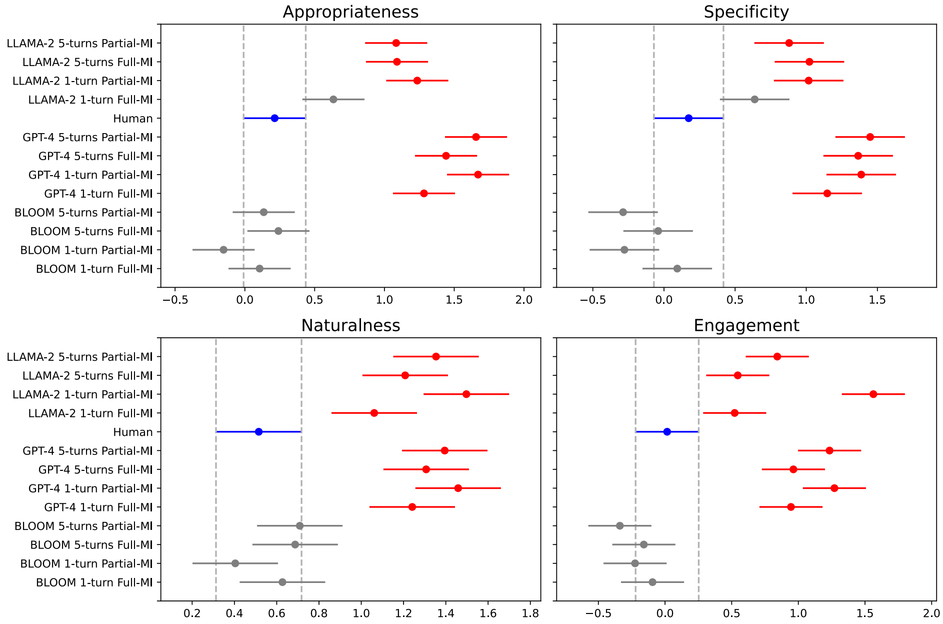


Figure 2.C.1: The mean scores for each model were calculated using Tukey's HSD test. Dashed lines mark the boundaries of the human reflections' results (blue bars). Bars that exceed these lines show a significant difference (red bars) while overlapping (gray) bars suggest no significant difference from the human utterances. The significance level was set to 0.05 for this visualization.

2.D Paired Samples T-tests of Prompting Strategies

Dimension	5-turns vs 1-turn	Partial-MI vs Full-MI
Appropriateness	$t = 2.363, p = .018, (*p < .05)$	$t = 2.263, p = .024, (*p < .05)$
Specificity	$t = 0.965, p = .335, (p > .05)$	$t = -0.146, p = 0.884, (p > .05)$
Naturalness	$t = 1.191, p = .234, (p > .05)$	$t = 2.204, p = .028, (*p < .05)$
Engagement	$t = -2.313, p = .021, (*p < .05)$	$t = 4.205, p = .000, (*p < .05)$

Table 2.D.1: Multiple paired samples t-tests calculated across the 4 criteria to measure the effects of changing the conversation context size and amount of details provided about motivational interviewing.

2.E Correlation between the Independent Evaluation Criteria

We calculated Pearson correlation coefficients to explore the linear relationships between each pair of the four criteria. All combinations showed a positive correlation; appropriateness vs specificity ($r(158) = 0.72, p < .001$), appropriateness vs naturalness ($r(158) = 0.64, p < .001$), appropriateness vs engagement ($r(158) = 0.70, p < .001$), specificity vs naturalness ($r(158) = 0.55, p < .001$), specificity vs engagement ($r(158) = 0.69, p < .001$), naturalness vs engagement ($r(158) = 0.62, p < .001$). The correlation is visualized in Figure 6.3.

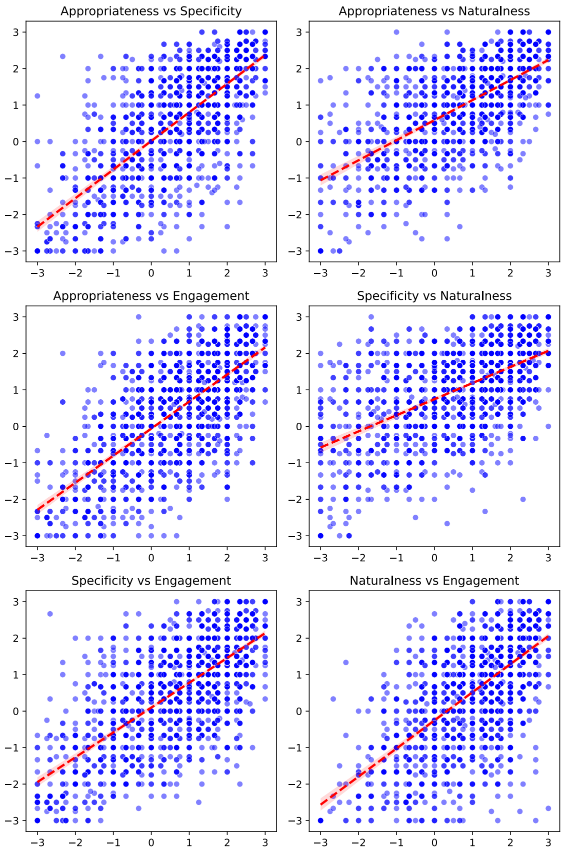


Figure 2.E.1: Scatter plots show Pearson ' s correlation (r) between all pairs of the four human evaluation criteria: appropriateness, specificity, naturalness, and engagement. Red lines indicate linear trends. All correlations are positive and significant (p < .001), with the highest between appropriateness and specificity and the lowest between specificity and naturalness.

2.F Prompt Template

Prompt Components	Content
Conversation context	**[The context of the conversation]** Conversation context (1-turn or 5-turn): Therapist: Yes, those were not really your moments, they were not really your smoking moments, that was a bit literally and figuratively, especially at the end of the day. [...] Client: Yes. Therapist: Yes, okay, so you say I am actually satisfied with the current state of affairs and ... Client: Yes I, I already said that, I like that with losing weight, I have a striving that I am between 85 and 90, that I still want to throw smoking out all the way, it is better anyway And cheaper.
Next MISC strategy (only for Full-MI setting)	**[The next MISC strategy for the therapist]** The next MISC strategy is: “Reflection”
MISC manual (only for Full-MI setting)	**[The descriptions of MISC strategy]** The definition of the MISC strategy: ’reflection’: reflection is a statement made by the therapist that captures and mirrors back the essence of what the client has said or expressed. [...] ’question’: question is made by the therapist to gain more clarity or to explore the client’s perspective, feelings, thoughts, or experiences. [...] ’therapist_input’: therapist_input is any other therapist utterance that is not codable as ’question’ or ’reflection’. [...]
MISC examples (only for Full-MI setting)	**[Two examples for each MISC code]** Example dialogues of each MISC code: ’reflection’: Example 1: Client: ’I’m scared of the consequences if I don’t stop smoking.’ Therapist: ’You’re expressing fear about the potential effects of continued smoking.’ [...] ’question’: Example 1: Client: ’I think I need to stop smoking.’ Therapist: ’Have you tried quitting before?’ [...]
Task instruction	**[The base instructions to explain the generation task]** Task: As a therapist of Motivational Interviewing, please generate the next appropriate utterance based on the conversation context. Restriction: you MUST NEVER ask new questions. The next therapist’s utterance is:

Table 2.F.1: The complete prompt template for MI reflection generation. The components involving MISC were added to the prompt only for the Full-MI prompt setting.

2.G Complete Contexts of Conversation Examples

Conversation Context
Therapist: Okay. Okay.
Client: Over a long– over a period of time.
Therapist: Mm-hmm. Okay. So let me just show you on this, um, card here, um, from a medical perspective, we know that there's good evidence that shows that women you're age who drink more than, um, three drinks on any one occasion at any- at any sitting or more than seven in a week, are likely to run into health problems from their drinking, either problems they're having right now or maybe in the future.
Client: Are you serious?
Therapist: Yeah. Surprised, ain't you?
Client: Seven drinks in a week doesn't seem like that much?
Therapist: It sounds like you're drinking a bit more than that right now.
Client: Yeah, and I-I didn't think it was much, but that makes it look like I'm drinking way too much.
Therapist: Mm-hmm. From this, uh, drinker's pyramid over on this side of the card, you can see that the-the score on your screening sheet shows that you're up in the harmful category of drinkers. So that it's likely you're drinking, you know, in quantities probably more than about 90% of the folks who come in to see me here at the clinic.
Client: Well, I know my drinking has increased a little since I got divorced and moved to [location], but I-I don't think it's that much of a problem.
Reflections
Therapist: It doesn't feel like a problem to you right now, yeah.
BLOOM 1-turn Full-MI: So, you know, you think that your drinking has increased a little since you got divorced and moved to [location], but you don't think it's that much of a problem.
BLOOM 1-turn Partial-MI: Well, I think, you know, it's important to look at the numbers. And, you know, I think, you know, the numbers are pretty clear that, you know, you are drinking more than you probably should be.
BLOOM 5-turns Full-MI: It sounds like you're concerned about your drinking habits at weekends.
BLOOM 5-turns Partial-MI: Well, I think that, um, you know, you might be right.
GPT-4 1-turn Full-MI: It appears that you're acknowledging your increased drinking, although you're not quite seeing it as a significant issue yet.
GPT-4 1-turn Partial-MI: I understand. it seems like you've gone through some significant life changes recently with your divorce and move to [location], which might have impacted your drinking habits. acknowledging this is an essential first step. just remember, it's not about labeling yourself as a problem drinker, it's more about understanding the relationship you have with alcohol and how it might be affecting your life.
GPT-4 5-turns Full-MI: So you believe your drinking isn't a big issue despite it has increased a bit since your divorce and move.
GPT-4 5-turns Partial-MI: It sounds like you have been through a lot with the divorce and moving to a new city. it's completely normal to try and cope with these life changes. acknowledging this is already a big step in the right direction. this doesn't mean your drinking is necessarily a problem, but it's worth keeping an eye on to ensure it doesn't lead to potential health issues.
Llama-2 1-turn Full-MI: It sounds like you're downplaying the significance of your drinking habits.
Llama-2 1-turn Partial-MI: I see. Well, let's take a closer look at this drinker's pyramid. It's true that you're not at the very top of the pyramid, but you're still in the harmful range. And it's interesting that you mention your recent divorce and move to [location]. Can you tell me a little more about how your drinking has changed since those events?
Llama-2 5-turns Full-MI: It sounds like you're acknowledging that your drinking has increased, but you're not sure it's a problem.
Llama-2 5-turns Partial-MI: Mm-hmm. It sounds like you're not sure if your drinking is a problem or not.

Table 2.G.1: The full conversation context of the example given in Table 2.4 which is discussed in Section 2.6.

Conversation Context
Therapist: Okay.
Client: So-
Therapist: Got it.
Client: -it ties in with the whole belonging thing.
Therapist: Okay. Well, when you think about your drinking and, sort of, looking forward in 5 years or 10 years, how do you think that drinking might fit in with your life?
Client: Um, it probably will have a very small role, if any-
Therapist: Mm-hmm.
Client: -in my life, considering I have big plans for my future. So.
Therapist: Okay. Can you expand on that a little bit more? What do you mean?
Client: I mean, I don't think I would be, you know, drinking as much or going to parties as much if I'm studying for school, or in college, or trying to get in law school and things like that.
Reflections
Therapist: Okay, so, the-the alcohol is something that might make it harder to reach your goals?
BL 1-turn Full-MI: Sounds like you're concerned about your drinking habits at weekends. And you also seem to be concerned about the consequences of your drinking if you don't stop.
BL 1-turn Partial-MI: Okay. So, you think that if you were in college, you would be less likely to drink?
BL 5-turns Full-MI: So, you think that drinking might be a problem for you in the future, but you don't think it's a problem now.
BL 5-turns Partial-MI: Okay. So, you think that drinking would be less important in your life if you were doing other things.
G4 1-turn Full-MI: So, in your view, pursuing education and focusing on your career would naturally steer you away from heavy drinking and partying. it's like you see these goals as your pathway to change.
G4 1-turn Partial-MI: It sounds like you're acknowledging that your current behaviors could be a hindrance to achieving your academic goals. recognizing this is a crucial step towards making changes. you seem to have a good understanding of the relationship between your actions and the outcomes you desire.
G4 5-turns Full-MI: So you're envisioning a future where your priorities have shifted more towards your studies and career goals, leaving less room for drinking and partying.
G4 5-turns Partial-MI: I really admire your commitment towards your future goals and your understanding about how certain behaviors could distract from those goals. you seem to have a really clear vision for your future which is truly inspiring.
L2 1-turn Full-MI: It sounds like you're suggesting that your drinking and partying behavior may be linked to your current academic and career goals, and that pursuing those goals might lead to a reduction in that behavior. Is that right?
L2 1-turn Partial-MI: So, it sounds like you're saying that your substance use and party lifestyle are a way of coping with the stress and pressure of school and your future plans. Is that right?
L2 5-turns Full-MI: It sounds like you're envisioning a future where your priorities shift, with less time for partying and more focus on your academic and professional pursuits. Is that right?
L2 5-turns Partial-MI: Okay. So it sounds like you're thinking about how your drinking might impact your future goals and plans. That's a really important thing to consider. Can you tell me more about what you hope to achieve in the next 5 or 10 years? What are some of those big plans that you have?

Table 2.G.2: The full conversation context of the example given in Table 2.5 which is discussed in Section 2.6.

2.H TrueSkill Values

Model	**Turns**	**MI**	**Mean (μ)**	**S.D. (σ)**
Human	n.a.	n.a.	26.98	0.85
GPT-4	5-turns	Full-MI	24.73	0.83
GPT-4	5-turns	Partial-MI	**29.60**	0.90
GPT-4	1-turn	Full-MI	21.22	0.81
GPT-4	1-turn	Partial-MI	26.72	0.85
Llama-2	5-turns	Full-MI	22.30	0.81
Llama-2	5-turns	Partial-MI	24.80	0.82
Llama-2	1-turn	Full-MI	19.38	0.82
Llama-2	1-turn	Partial-MI	24.34	0.82
BLOOM	5-turns	Full-MI	15.68	0.86
BLOOM	5-turns	Partial-MI	16.74	0.85
BLOOM	1-turn	Full-MI	16.18	0.85
BLOOM	1-turn	Partial-MI	16.89	0.86

Table 2.H.1: TrueSkill mean rating values (μ) and standard deviations (σ) for each model, conversation context size, and MI strategy combination.

3

Creation of Motivational Interviewing (MI) Dataset

This chapter is based on the following publication:

Authors: Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos TP Dobber, Jos A. Bosch.

Original title: Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study

Published in: International Conference on Computational Linguistics (COLING 2024)

Abstract

Behavioral coding (BC) in motivational interviewing (MI) holds great potential for enhancing the efficacy of MI counseling. However, manual coding is labor-intensive, and automation efforts are hindered by the lack of data due to the privacy of psychotherapy. To address these challenges, we introduced BiMISC, a bilingual dataset of MI conversations in English and Dutch, sourced from real counseling sessions. Expert annotations in BiMISC adhere strictly to the motivational interviewing skills code (MISC) scheme, offering a pivotal resource for MI research. Additionally, we presented a novel approach to elicit the MISC expertise from Large language models (LLMs) for MI coding. Through the in-depth analysis of BiMISC and the evaluation of our proposed approach, we demonstrated that the LLM-based approach yields results closely aligned with expert annotations and maintains consistent performance across different languages. Our contributions not only furnish the MI community with a valuable bilingual dataset but also spotlight the potential of LLMs in MI coding, laying the foundation for future MI research.

3.1 Introduction

Motivational Interviewing (MI) is an essential, directive, client-centered counseling technique, that aims to elicit clients' behavioral change [117]. It can boost intrinsic motivation and collaboration between therapists and clients by effectively addressing ambivalence and enhancing self-efficacy [118], improving clients' adherence to the therapists' interventions [19]. Without the use of MI, traditional techniques can potentially cause resistance and disengagement from clients due to their confrontational, paternalistic ways of thinking [117].

Behavioral coding (BC) is the practice of systematically observing and categorizing the behaviors of therapists and clients [117–119]. Output codes can provide valuable insights for professional practitioners (e.g., therapists and counselors), regarding behavioral patterns and their connections to therapeutic outcomes.

However, conducting behavioral coding in MI is challenging, as it relies on domain expertise [120, 121] and large-scale datasets with human annotations [122–124]. Recent research has demonstrated the effectiveness of natural language processing (NLP) approaches in supporting behavioral coding. Early efforts primarily utilized statistical models such as N-grams [125] and Conditional Random Fields (CRFs) [126]. More recent work has shifted towards deep learning methods, including Recurrent Neural Networks (RNNs) [119, 127], Convolutional Neural Networks (CNNs) [128], and Bidirectional Gated Recurrent Units (BiGRUs) with attention mechanism [122].

Client Utterance	Annotation	
	BiMISC	AnnoMI
I was helped.	FN	NT
But I just have to keep it up because I still have to use medicines for a month and a half.	R+	
It is a kind of course, right?	ASK	

Table 3.1: An example of annotation of a client utterance in the BiMISC and AnnoMI datasets. BiMISC uses more fine-grained codes, e.g., FN (Follow Neutral), R+ (Reflection Listening), ASK (Asking Question), whereas AnnoMI uses broader labels such as NT (Neutral Talk). Besides, the use of multiple codes in BiMISC allows for a multi-perspective and in-depth understanding of the client ' s intentions. These are non-trivial insights for therapists conducting MI treatment in psychotherapy.

These aforementioned approaches require large-scale high-quality data and computing resources. An even greater challenge is that the majority of MI resources are not publicly accessible, primarily due to privacy concerns. For example, [129] introduces a dataset with 277 conversations covering 10 MI codes, however, the dataset is no longer accessible. To the best of our knowledge, there are only two publicly available datasets, namely, AnnoMI [130] and motivational interviewing treatment integrity (MITI) [131]. However, there are still challenges: (1) they do not consist of conversations from real MI counseling sessions; (2) AnnoMI comprises only six coarse-grained MISC codes, while MITI solely includes codes for therapist behaviors and lacks codes for the client's actions; (3) They assign only one code to each utterance, which may not fully capture the complex intentions behind it.

In this work, we introduced BiMISC, a bilingual dataset of client and therapist utterances in English and Dutch, annotated with behavioral codes following MISC scheme: (1) BiMISC consists of conversations collected from real MI counseling sessions in psychotherapy; (2) BiMISC comes with fine-grained behavioral codes strictly grounded on the MISC scheme; (3) BiMISC features multiple codes instead of a single code for each utterance.

Large language models (LLMs) have been proven to be effective in both providing accurate responses in open-domain [132] and eliciting expertise in various domain-specific applications, e.g., medical treatment [133], legal judgment analysis [134] and qualitative data analysis [135–137]. In this work, we aimed to explore the potential of utilizing the MISC scheme with LLMs to directly generate MISC codes that closely align with expert annotations. Specifically, we leveraged MISC codes and their definitions in the MISC manual to elicit expertise from an LLM for MI coding. We first designed a prompt template, including task instruction, MISC manual, MISC examples, and historical conversations. Next, we randomly selected approximately 3% of the data to serve as test samples for conducting trial experiments on MISC coding. We continuously refined the prompt manually until the output codes are in alignment with human annotations. Last, we conducted experiments and evaluations on the full dataset.

Our contributions can be summarized as follows:

- We collected and released BiMISC, the first bilingual MI dataset in both English and Dutch with expert-annotated MISC codes;
- We proposed a MISC coding approach by eliciting MISC expertise from LLM;
- We conducted extensive experiments and analysis of the proposed approach, demonstrating its effectiveness in MI behavioral coding.

3.2 Related Work

3.2.1 Motivational Interviewing

Motivational Interviewing (MI) [117] is a counseling technique aimed at boosting an individual's motivation to make behavioral changes. It addresses doubts or ambivalence about change and strengthens a person's belief in their ability to make positive changes [118]. By fostering a supportive environment, MI helps individuals find their own reasons to change and has shown success in areas like health promotion and substance abuse [19].

In Motivational Interviewing (MI), behavioral coding (BC) is important, serving as a means to observe and categorize behaviors demonstrated by therapists and clients during MI counseling sessions [117, 119]. This systematic categorization provides therapists with insightful perspectives and facilitates steering of therapeutic interventions more efficiently. Behavioral coding (BC) hinges on a defined scheme consisting of predetermined codes, each associated with specific MI-associated behaviors [118, 120]. Once established, these codes can be systematically assigned to the transcripts of MI counselings. Behavioral coding (BC) empowers researchers to discern MI behavioral patterns and link them with therapeutic outcomes, deepening our understanding of the MI intervention process.

To this end, the MI research community has developed validated coding schemes for behavioral coding in MI. Notable among these are Motivational Interviewing Skills Code (MISC) [103, 120] and Motivational Interviewing Treatment Integrity (MITI) [138]. A key difference between MISC and the MITI is their focus: while MISC offers behavioral codes for both therapists and clients, MITI predominantly concentrates on therapist behaviors. These comprehensive coding schemes assess MI-specific behaviors manifested in therapist-client interactions, such as the use of questions and reflections. They have been widely employed in MI research for various purposes, including assessing therapist adherence, measuring the effectiveness of MI training, and examining the relationship between specific MI behaviors and therapeutic outcomes. In our research, we opted for the MISC coding scheme [103] to annotate behaviors of both therapist and client.

3.2.2 MI datasets

Resources for MI are limited due to the sensitive nature of the topics discussed in counseling and psychotherapy. For instance, psychotherapy transcripts from platforms like Alexander Street [139] are not publicly accessible because of privacy. While annotated MI datasets exist, such as the collection of MI conversational recordings by [129], these data are not publicly accessible. To the best of our knowledge, there are two publicly available MI datasets. The first one is AnnoMI [130], a dataset compiled from automatic transcriptions of MI recordings from video-sharing platforms. This dataset is annotated with MI codes based on a self-constructed coding scheme (which is a subset/regroup of MISC). The second one MI dataset [131] comprises dialogues from social forums. These dialogues are annotated based on the MITI [138] coding scheme by crowdsourcing annotators. In this work, we introduced BiMISC, which is a bilingual dataset available in both English and Dutch. BiMISC comprises conversations sourced directly from actual MI counseling sessions in psychotherapy. And BiMISC was annotated strictly grounded on the MISC scheme [103] by MI experts.

3.2.3 MI coding approaches

The field of MI has benefited significantly from established coding schemes like MISC. However, the manual coding process associated with these schemes is labor-intensive, necessitating specialized training and expertise [120,121]. This has resulted in a growing demand for efficient methods, paving the way for the development of automatic MI coding approaches. Initial approaches in this direction lean on statistical models, with prior work exploring the utility of N-grams [125], topic models [121], and CRFs [126] for MI coding. With advancements in computational power, the focus has shifted towards deep learning models. Recent work has studied the applications of RNNs [119, 127], CNNs [128], and BiGRUs with attention mechanisms [122]. While these models show promise, they also bring their own challenges, especially the need for substantial data. A primary barrier to the wider adoption of automatic MI coding is the limited access to MI resources, due to privacy concerns within psychotherapy.

Most recently, large language models (LLMs) have demonstrated effectiveness in providing accurate open-domain responses [132] and in showcasing expertise across various domain-specific applications, such as medical treatment [133,140], legal judgment analysis [134], and qualitative data analysis [136,137,141], especially in zero-shot scenarios [142,143]. Given their capabilities, LLM offers potential for MI coding, which could alleviate the need for extensive training data required by previous research [119, 122,127,128]. Therefore, we explored the feasibility of eliciting domain expertise of the MISC scheme from LLMs for efficient MISC coding.

3.3 Dataset Creation

Figure 3.1: Process of the construction of the bilingual dataset: BiMISC.

In this section, we outline the creation of the BiMISC dataset as shown in Figure 3.1. First, we collected raw conversations between therapists and clients from MI counseling sessions (§ 3.3.1). Second, we introduced the MISC scheme for human annotation (§ 3.3.2). Last, we reported the statistics of the proposed BiMISC dataset (§ 3.3.3).

3

Therapist Code	Description (abbreviated version)	Example
Open question (OQ)	Asking questions for a wide range of answers.	Can you tell me more about your drinking habits?
Closed question (CQ)	Asking questions for concise answers: “Yes” or “no”, a number.	Did you use heroin this week?
Simple reflection (SR)	Conveying shallow understanding without additional information.	You don’t want to do that.
Complex reflection (CR)	Conveying deep understanding with additional information.	That’s where you drew the line.
Advice (ADV)	Providing suggestions or recommendations.	Consider starting with small, manageable changes like taking a short walk daily.
Affirm (AFF)	Conveying positive or complimentary information.	You did well by seeking help.
Direct (DIR)	Offering an imperative order, command, or direction.	You’ve got to stop drinking.
Emphasize control (EC)	Emphasizing client’s freedom of choice.	It’s up to you to decide whether to drink.
Facilitate (FA)	Encouraging the client to keep sharing.	Tell me more about that.
Filler (FIL)	Fitlering utterances are not related to behavior change.	Good Morning!
Giving information (GI)	Offering relevant information, explanations, or feedback.	There are several treatment options available for managing stress.
Support (SP)	Offering encouragement and reassurance	I’m here to support you through your recovery journey.
Structure (STR)	Offering a treatment process during the client’s journey.	First, let’s discuss your drinking, and then we can explore other issues.
Warn (WAR)	Offering a warning or negative consequences.	You could go blind if you don’t manage your blood sugar levels.
Permission seeking (PS)	Asking for consent before providing information or advice.	May I suggest a few stress management techniques?
Opinion (OP)	Expressing a viewpoint or judgment	In my opinion, addressing your stress can help reduce your drinking.

Client Code	Description	Example
Follow/Neutral (FN)	No indication of client inclination toward or away from change.	Yeah.
Ask (ASK)	Asking for clarification or information.	What treatment options are available?
Commitment (CM+/CM-)	An agreement, intention, or obligation regarding future change.	I will try to reduce my drinking.
Taking step (TS+/TS-)	Concrete steps the client has recently taken to make a change.	I threw away all of my cigarettes.
Reason (R+/R-)	Rationale, basis, justification, or motive to make a change.	It would be so good for my kids.
Other (O+/O-)	Other statements clearly reflect intention of change.	My family doesn’t believe I can quit.

Table 3.2: MISC codes in the BiMISC dataset. The symbols ”+” and ”-” represent the client’s desire to change (+) or not change (-) their behaviors with CM, TS, R or O intention.

Role	AnnoMI Codes	BiMISC Codes
Therapist	Question (QS)	OQ, CQ
	Reflection (RF)	SR, CR
	Therapist Input (TI)	ADV, AFF, DIR, EC, FA, FIL, GI, SP, STR, WAR, PS, OP
Client	Neutral Talk (NT)	FN, ASK
	Change Talk (CT)	CM+, TS+, R+, O+
	Sustain Talk (ST)	CM-, TS-, R-, O-

Table 3.3: Mapping relationship between the codes in BiMISC (fine-grained) and AnnoMI (coarse-grained).

3

3.3.1 Raw data collection

Initially, we collected 80 audio recordings of conversations between 18 clients and therapists in real MI counseling sessions, conducted in Dutch. The therapists transcribed these recordings, each averaging 108 utterances per conversation. During transcription, they also corrected typos and grammatical errors and anonymized sensitive information (e.g., names and addresses). Next, we utilized a machine translator [144] to translate the Dutch transcripts into English, followed by post-editing by two Dutch Master's students to improve translation quality. Compared to AnnoMI, our raw data contains a similar number of utterances (8,572 vs. 8,839). Each conversation in our dataset is longer, with significantly more turns on average (108 vs. 80 utterances per conversation). Moreover, our conversations were drawn from real counseling sessions, ensuring both authenticity and relevance.

3.3.2 MISC annotation

We followed the MISC 2.1 scheme [103][1] and defined an annotation manual that contains MISC codes with their descriptions and examples (See Table 3.2). The certified MI therapists from the Dutch institute who conducted the MI counseling and initially recorded the conversation assigned each utterance with the appropriate MISC codes. Each utterance was coded to reflect all applicable MISC behaviors, leading to a multi-code annotation. For example, as shown in Table 3.1, for the utterance "I was helped. But I just have to keep it up because I still have to use medicines for a month and a half. It is a kind of course, right?", it should be assigned as "follow/neutral," "reason+," and "ask" rather than just a single code. In addition, the therapists annotated the most precise fine-grained MISC codes, as shown in Table 3.2. To ensure the quality of the dataset and its annotations, we provided annotators with a comprehensive guideline. They were encouraged to flag and discuss ambiguous cases. Furthermore,

[1] https://digitalcommons.montclair.edu/cgi/viewcontent.cgi?article=1026&context=psychology-facpubs

Dataset	AnnoMI	BiMISC
# Utterances	8,839	8,572
# Conversations	110	80
# Avg utterances / conversation	80	108
# MISC codes	6	26
# Therapist codes	3	16
# Client codes	3	10
Language	English	Dutch & English
Multiple codes / utterance	False	True

Table 3.4: Comparison of the AnnoMI dataset and the proposed BiMISC dataset. Note that BiMISC provides fine-grained multiple codes for each utterance and bilingual, parallel, in-depth conversations. The use of fine-grained multiple codes provides therapists with profound insights for conducting MI treatment in psychotherapy.

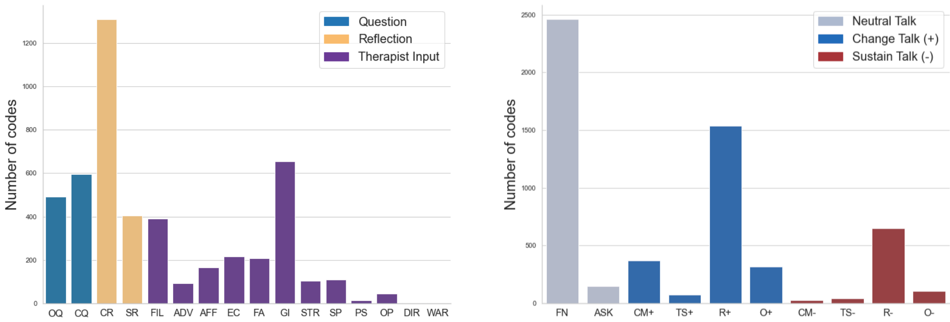


Figure 3.2: Distribution of fine-grained codes for the client (right) and therapist (left) in BiMISC.

a subset of the data underwent double-annotation to assess inter-annotator agreement, ensuring the reliability and consistency of the annotations.

3.3.3 Data statistics

Table 3.4 compares the statistics of the proposed BiMISC dataset and AnnoMI dataset.

The AnnoMI dataset is a publicly available MI dataset consisting of 110 conversations with 8,839 utterances. It only partially introduced 6 *coarse-grained* MISC codes, including 3 therapists' codes (i.e., question, reflection, and therapist input), and 3 clients' codes (i.e., neutral talk, change talk, and sustain talk).

BiMISC, on the other hand, consists of 80 conversations with 8,572 utterances in both Dutch and English. These utterances are annotated with a total of 26 *fine-grained* behavioral codes derived from the MISC scheme, with 16 codes corresponding to therapist behaviors and 10 attributed to client behaviors (See Table 3.2).

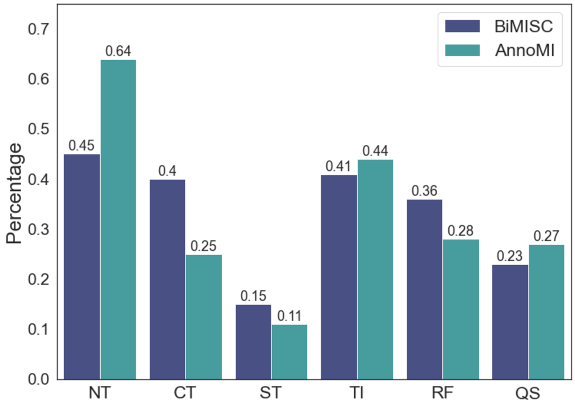


Figure 3.3: Distribution of coarse-grained codes in AnnoMI dataset and BiMISC dataset.

Table 3.3 shows the mapping relationships between fine-grained codes (used in BiMISC) and coarse-grained codes (used in AnnoMI). For example, the fine-grained codes “open question (OQ)” and “closed question (OQ)” can be mapped to the coarse-grained code “question (QS)”.

Figure 3.2 shows the distribution of fine-grained codes for therapists (left) and clients (right) in BiMISC. To make the two datasets comparable, we consider the coarse-grained MISC does in both BiMISC and AnnoMI, and plot the distribution of the codes in Figure 3.3. We see that BiMISC has a more balanced distribution of the codes compared with AnnoMI.

3.4 Experimental Setup

3.4.1 Research questions

We seek answers to the following research questions by the experiments:

(RQ1) Is the use of fine-grained codes advantageous for LLMs in predicting MISC codes?

(RQ2) What are the key factors that affect the elicitation of MISC expertise from LLMs for MISC codes?

(RQ3) Do LLMs maintain consistent performance in MISC code prediction across different languages?

3.4.2 Task definition and evaluation

We defined the MISC coding task as the classification of therapist or client utterances into specific MISC codes. Utilizing an LLM, we provided the model input with a prompt, including a task instruction, the MISC manual, MISC examples, and the historical conversations. And LLM subsequently generates MISC codes as an output.

We conducted evaluations as a classification task using the following metrics:

- Accuracy: the fraction of responses that have been categorized into a correct code out of all responses.
- Precision: measures the percentage of codes identified as positive that are actually positive.
- Recall: measures the percentage of actual positive codes that were identified correctly.
- Macro F1 [145]: provides a well-rounded metric that factors in both precision and recall. This is important given the imbalanced distribution of codes in MI conversations.

3.4.3 Benchmark models

We employed three prominent LLMs as benchmarks, including two commercial and one open-source. We set the hyper-parameter temperature as `0` to control the randomness of generation and ensure reproducibility.

GPT-3.5 We selected `gpt-3.5-turbo` [2] as a commercial LLM benchmark. It has been optimized for better alignment with human instructions and chat interactions.

GPT-4 We selected `gpt-4` [3] as another commercial LLM benchmark. It demonstrates outstanding performance in providing accurate responses as human instruction, notably in zero-shot scenarios.

Flan-T5 We selected `flan-t5-xxl` [4] as an open-source LLM benchmark, known for its advanced capabilities in various NLP tasks [146], optimized for better alignment with human instructions.

We also explored `Llama-2-13b-chat-hf` [5] as an open-source LLM benchmark. However, it struggled to differentiate between the various MISC codes, often leading to the generation of unintended outputs that deviate from the given prompt.

3.4.4 Prompt template design

We elaborated a comprehensive prompt template by the following steps:

(1) We crafted an initial prompt template, containing task instructions, MISC guidelines, illustrative examples, and historical conversations;

[2] `https://platform.openai.com/docs/models/gpt-3-5`

[3] `https://platform.openai.com/docs/models/gpt-4`

[4] `https://huggingface.co/google/flan-t5-xxl`

[5] `https://huggingface.co/meta-llama/Llama-2-13b-chat-hf`

(2) We conducted meticulous manual verification and refinement until the resulting output codes meet our predefined expectations.

An example of the prompt template is detailed in Appendix 3.A. Notably, there are two integral components within the prompts:

MISC CODING MANUAL: We introduced the definition of a list of role-specific codes with their names and comprehensive descriptions (see Table 3.2). The role of the current speaker is either therapist or client. The descriptions are carefully crafted and condensed into a scheme handbook [103][1] by experts specializing in MI in psychotherapy.

MISC CODING EXAMPLES: We offered each code two examples of client-therapist utterance pairs. These examples were selected from the MISC scheme handbook [1].

3.5 Outcomes

3.5.1 Overall performance (RQ1)

Dataset	Model	Macro F1	Client's Codes			Therapist's Codes		
		All	NT	CT	ST	QS	RF	TI
AnnoMI	GPT-3.5	0.53	0.69	0.56	0.36	0.63	0.54	0.39
	Flan-T5	0.60	**0.79**	0.52	0.29	0.81	0.58	0.62
	GPT-4	**0.73**	0.76	**0.69**	**0.46**	**0.84**	**0.74**	**0.87**
BiMISC	GPT-4	0.68	**0.70**	0.73	0.42	0.83	0.65	0.75
	GPT-4 + mapping	*0.68*	0.44	*0.73*	**0.55**	**0.86**	**0.70**	**0.80**

Table 3.5: MISC coding performance, evaluating coarse-grained codes using the Macro F1 score for clients and therapists. We conduct single-code classification (AnnoMI) and multi-code classification (BiMISC). The "mapping" indicates that we conduct fine-grained multi-code classification and then map the fine-grained codes to coarse-grained codes following Table 3.3.

To address RQ1, we conducted experiments on AnnoMI and BiMISC datasets respectively, and evaluated the performance on coarse-grained codes and fine-grained codes. Table 3.5 shows MISC coding performance on AnnoMI and BiMISC, evaluated by F1 score on coarse-grained.

First, fine-grained codes can better elicit MISC expertise from LLMs for MISC coding. In the BiMISC dataset, GPT-4 + mapping (predicting fine-grained codes and mapping them into coarse-grained codes) achieved substantial improvements or showed comparable results when compared to GPT-4 (predicting coarse-grained codes directly). Specifically, F1 scores on ST, QS, RF, and TI increased 13%, 3%, 5%, 5%, respectively. This is because fine-grained multiple codes are mutually beneficial:

fine-grained multiple codes enable a comprehensive and multi-dimensional expression of the client's intentions. These insights held significant value for therapists in delivering effective MI treatment within the realm of psychotherapy. The only exception is on NT. Analyzing Figure 3.4, we observed that the fine-grained classification (b) has a higher false negative rate and a lower true positive rate for the NT code compared to the coarse-grained classification (a). This is due to the fact that the definition of NT is not as clear-cut as CT and ST and it is frequently disregarded by LLMs in multi-code scenarios.

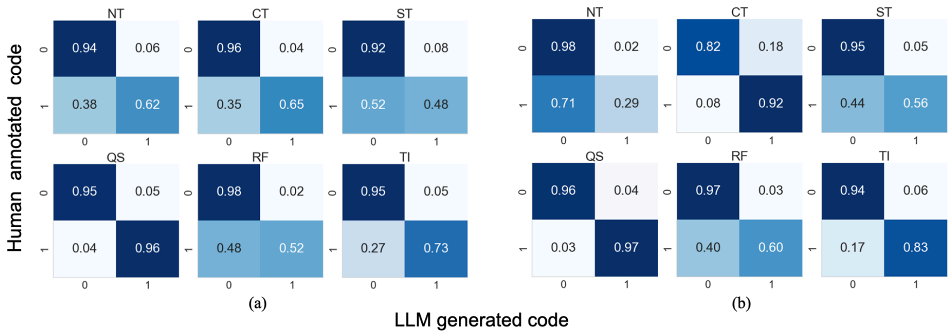


Figure 3.4: Accuracy of multi-code classification on the BiMISC dataset by GPT-4, using coarse-grained codes (a) and fine-grained codes (b).

Second, GPT-4 exhibited the highest performance, with Flan-T5 and GPT-3.5 following in the evaluation of single-code classification on AnnoMI. Notably, Flan-T5 achieved the best performance for the NT code. Specifically, GPT-4 significantly outperformed Flan-T5 and GPT3.5 by 13% and 20% in terms of overall performance. So we conducted thorough analytical studies utilizing GPT-4 as the chosen LLM.

Third, multi-code classification (See Figure 3.4) was generally more challenging than single-code classification (See Figure 3.5). The true positive prediction accuracy for multi-code classification was typically quite high, but it is important to note that false negatives can be relatively high in certain scenarios. For example, GPT-4+mapping achieved 0.71, 0.44, and 0.40 for NT, ST and RF.

3.5.2 Elicitation of MISC expertise (RQ2)

To address RQ2, we conducted an ablation study to assess how two key factors (i.e., MISC manual and examples) influence the elicitation of MISC expertise from LLM, as shown in Table 3.6.

First, the choice of prompt substantially influenced the LLM's performance.

[5]This sampled data includes all six behaviors and has a code distribution similar to the entire AnnoMI dataset. The costs are approximately \$5 and \$15 per 1,000 codes for GPT-3.5 and GPT-4, respectively.

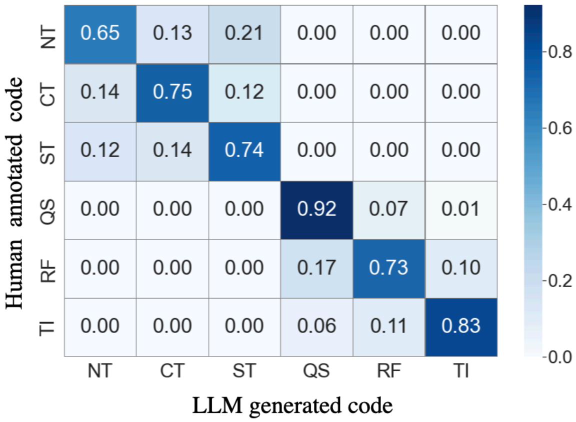


Figure 3.5: Accuracy of single-code classification on the AnnoMI dataset as performed by GPT-4.

Model	MISC Manual	# Examples	Macro F1	Client's codes			Therapist's codes		
			All	NT	CT	ST	QS	RF	TI
GPT-3.5	True	0	0.55	0.74	0.56	0.27	0.70	0.57	0.44
	True	2	0.55	0.71	0.54	0.33	0.58	0.56	0.57
	False	0	0.38	0.75	0.53	0.22	0.16	0.47	0.15
	False	2	0.45	0.72	0.50	0.30	0.17	0.30	0.17
Flan-T5	True	0	0.62	0.81	0.55	0.26	0.85	0.61	0.64
	True	2	0.45	0.72	0.50	0.30	0.17	0.30	0.17
	False	0	0.52	0.74	0.28	0.12	0.81	0.49	0.69
	False	2	0.53	0.82	0.53	0.05	0.85	0.30	0.64
GPT-4	True	0	**0.73**	**0.91**	**0.67**	0.40	**0.87**	**0.80**	0.74
	True	2	0.67	0.77	0.56	**0.43**	0.75	0.67	**0.83**
	False	0	0.58	0.72	0.45	0.34	0.73	0.60	0.67
	False	2	0.58	0.87	0.60	0.31	0.83	0.24	0.67

Table 3.6: The performance of benchmark models on the sampled 15% AnnoMI with equal distribution of entire dataset, evaluated using Macro F1 score, with consideration given to various prompt setups.

Specifically, in GPT-4 when using the MISC manual, the F1 scores increased from 0.58 to 0.73, marking a 15% improvement. This confirmed that the MISC manual can elicit MISC expertise from LLM for classification. Second, examples are beneficial for LLMs, particularly when MISC manuals are not available. The macro F1 score of GPT-3.5 increased by 7%, while Flan-T5 saw a 1% increase when two examples were used without the MISC manual. Third, from a model performance standpoint, GPT-4 led the pack, followed by Flan-T5, with GPT-3.5 coming in last. Consistent results across different prompt setups with these three models affirmed the generalizability of LLMs for MISC classification and support our findings.

3.5.3 Multi-lingual analysis (RQ3)

To address RQ3, we conducted a comparative analysis of fine-grained classification using both English and Dutch MI conversations on BiMISC. We kept our experimental setup consistent, ensuring that the prompt setup matches the language of the MI conversations being assessed.

The results, as displayed in Table 3.7, showed that GPT-4's performance remains comparable and consistent across both English and Dutch MI conversations, suggesting its ability to understand and classify MI conversations are not limited to English, which highlights GPT-4's robust multi-lingual capabilities. The multi-lingual consistency of GPT-4 in MI coding, irrespective of the language, indicated its potential as a valuable tool in multilingual psychotherapy contexts. Such a tool can assist therapists in diverse settings, ensuring that the nuances of MI conversations are accurately captured and analyzed. Our investigation into RQ3 provided affirmative evidence. Consistent multi-lingual performance in MISC classification paved the way for broader applications in multi-lingual psychotherapy contexts.

Therapist	**All**	QS		RF		TI											
		OQ	CQ	SR	CR	AFF	ADV	DIR	EC	FA	FIL	GI	SP	STR	WAR	PS	OP
English (EN)	.31	.71	.62	.28	.25	.32	.35	.00	.00	.12	.21	.59	.22	.15	.00	.50	.00
Dutch (NL)	.33	.70	.62	.29	.30	.39	.56	.00	.10	.10	.24	.66	.14	.15	.00	.40	.00

Client	**All**	NT		CT				ST			
		FN	ASK	CM+	TS+	R+	O+	CM-	TS-	R-	O-
English (EN)	.32	.35	.57	.35	.14	.55	.12	.25	.36	.42	.06
Dutch (NL)	.30	.47	.57	.29	.16	.53	.19	.18	.22	.43	.00

Table 3.7: The fine-grained classification in English (EN) and Dutch (NL) on the BiMISC dataset, evaluated using the Macro F1 score for therapist (upper) codes and client (lower) codes.

3.6 Limitations

While our work introduced a novel approach for MISC coding supported by the creation of a new MI dataset, we acknowledge the following limitations: First, the size of BiMISC is somewhat limited, potentially impacting the performance of fine-grained classifications, particularly concerning the underrepresented codes. One potential remedy is the utilization of data augmentation techniques to enhance the representation of these codes. Second, our evaluation of MISC coding highly depends on human annotations. This approach may introduce biases, leading to potential inaccuracies in the evaluation process. Incorporating a multi-annotator system complemented by cross-validation might help mitigate individual biases. Third, our multi-code classification relies heavily on a predefined confidence threshold for LLM. The LLM is instructed to give multiple codes only when its confidence exceeds this threshold, this can considerably affect the outcomes. Adaptive thresholding techniques could be explored to optimize multi-code classification.

3.7 Conclusion and Future work

We introduced BiMISC, a bilingual dataset comprising MI conversations in both English and Dutch. We built BiMISC using expert annotation, carefully aligned with the MISC scheme. Our comprehensive analysis and comparison with AnnoMI highlighted BiMISC's distinctiveness and novelty. Furthermore, the promising outcomes from our experiments not only spotlighted the potential of LLMs in MISC coding but also highlighted the unique characteristics and advantages of BiMISC.

In the future, we plan to study further the fine-grained classification of MI conversations, specifically addressing the challenges posed by imbalanced codes in MISC classification. Moreover, we envision leveraging the outcomes of MISC classification as directives for natural language generation within MI. This sets the stage for incorporating controllable natural language generation in sensitive domains like psychotherapy.

Ethics Statement

Data Anonymization

The data utilized in this work originates from real MI counseling sessions and thus contains sensitive information. We received consent from the client to allow us to make recordings of the MI counseling. To protect the privacy of the individuals involved, we implemented rigorous data anonymization procedures. All identifiable information, including names, addresses, and any specific personal details, were meticulously removed or replaced with pseudonyms to ensure confidentiality and anonymity.

Expert Annotation

To maintain the integrity and quality of the data, annotation was conducted by qualified experts in the field of Motivational Interviewing. These experts have significant experience and training in MI, ensuring that the annotation process is executed with a deep understanding of the therapy's nuances and ethical considerations. The experts are also bound by confidentiality agreements to safeguard the privacy of the individuals in the MI recordings and transcripts.

Ethical Concerns

We acknowledged and carefully considered the ethical implications throughout the research process. The work is strictly adherent to the ethical requirements of the institute. We also seek to minimize any potential harm or misuse of the information.

Chapter appendix

3.A An example of the prompt for classifying the MI code

Prompt	**[ROLE]: Therapist**
MISC manual	Definition of each code in MISC for [ROLE]: **[We give descriptions of each MISC code according to the [ROLE]]** ’reflection’: reflection is a statement made by the therapist that captures and mirrors back the essence of what the client has said or expressed. [...] ’question’: question is made by the therapist to gain more clarity or to explore the client’s perspective, feelings, thoughts, or experiences. [...] ’therapist_input’: therapist_input is any other therapist utterance that is not codable as ’question’ or ’reflection’. [...]
MISC examples	Examples of each code in MISC: **[We give TWO examples of each MISC code according to the [ROLE]]** ’reflection’: Example 1: Client: ’I’m scared of the consequences if I don’t stop smoking.’ Therapist: ’You’re expressing fear about the potential effects of continued smoking.’ [...] ’question’: Example 1: Client: ’I think I need to stop smoking.’ Therapist: ’Have you tried quitting before?’ [...] ’therapist_input’: Example 1: Client: ’I feel anxious lately.’ Therapist: ’Managing anxiety is possible with strategies like relaxation techniques and mindfulness.’ [...]
Historical conversations	Conversations: **[We give historical conversations and the utterance need to be classified]** Therapist: Yes, those were not really your moments, they were not really your smoking moments, that was a bit literally and figuratively, especially at the end of the day. [...] The utterance for classification: Therapist: Yes, and yes the weight does not go to me, but is that something that will be coming soon or you say that will only be next year.
Task instruction	Task: **[We give instruction to explain the MISC classification task]** Given the above Conversations, please identify the MISC codes for the last therapist’s last utterance. Provide the code based solely on these options: [’reflection’, ’question’, ’therapist_input’]. Provide only the selected codes without any additional text. Code is:

Table 3.A.1: The complete prompt template for MI reflection generation. The components involving MISC were added to the prompt only for the Full-MI prompt setting.

4

Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy

This chapter is based on the following publications:

Authors: Xin Sun, Jan de Wit, Zhuying Li, Jiahuan Pei, Abdallah El Ali, Jos A Bosch.

Original title: Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy

Published in: (Journal) Proceedings of the ACM on Human-Computer Interaction (Track Computer-Supported Cooperative Work And Social Computing CSCW)

Authors: Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, Jos Bosch.

Original title: Rethinking the Alignment of Psychotherapy Dialogue Generation with Motivational Interviewing Strategies

Published in: International Conference on Computational Linguistics (COLING 2025)

Abstract

Chatbots or conversational agents (CAs) are increasingly used to improve access to digital psychotherapy. Many current systems rely on rigid, rule-based designs, heavily dependent on expert-crafted dialogue scripts for guiding therapeutic conversations. Although advances in large language models (LLMs) offer potential for more flexible interactions, their lack of controllability and explanability poses challenges in high-stakes contexts like psychotherapy. To address this, we conducted two studies in this work to explore how aligning LLMs with expert-crafted scripts can enhance psychotherapeutic chatbot performance. In Study 1 (N=43), an online experiment with a within-subjects design, we compared rule-based, pure LLM, and LLMs aligned with expert scripts via fine-tuning and prompting. Results showed that aligned LLMs significantly outperformed the other types of chatbots in empathy, dialogue relevance, and adherence to motivational interviewing (MI) principles. Building on findings, we proposed "Script-Strategy Aligned Generation (SSAG)", a more flexible alignment approach that reduces reliance on fully scripted content while maintaining LLMs' therapeutic adherence and controllability. In a 10-day field Study 2 (N=21), SSAG achieved comparable therapeutic effectiveness to full-scripted LLMs while requiring less than 40% of expert-crafted dialogue content. These results demonstrated that expert alignment remains essential for LLM-powered psychotherapy, and that SSAG offers a more efficient path to develop controllable and therapeutically grounded chatbots. Beyond these results, this work advances LLM applications in psychotherapy by providing a controllable and scalable solution, reducing reliance on expert effort. By enabling domain experts to align LLMs through high-level strategies rather than full scripts, SSAG supports more efficient co-development and expands access to a broader context of psychotherapy.

4.1 Introduction

Chatbots, or conversational agents (CAs), are increasingly used in psychotherapy for behavioral interventions. These agents utilize evidence-based conversational techniques such as Motivational Interviewing (MI) [8] and Cognitive Behavioral Therapy (CBT) [20] to provide round-the-clock mental health support [4, 6, 7, 21, 147–149]. Traditionally, these agents have relied on rule-based approaches [12] grounded in expert-crafted scripts [13, 14] to ensure therapeutic adherence and safety. While effective, these approaches often produce rigid, non-adaptive conversations and require substantial expert effort to not only author dialogue content but also design the structured dialogue flows, as illustrated in Fig 4.1. Recent advances in Natural Language Generation (NLG) [150] have enabled more dynamic chatbot interactions. For example, MI chatbots have incorporated rephrasing and template-based generation [90] to improve engagement [4, 6, 73], while models trained on CBT data [21] demonstrated the potential to deliver structured and principle-based therapeutic dialogues. Hybrid approaches [151,152] have emerged to integrate LLMs with rule-based systems, with the aim of combining generative flexibility with expert-guided structure and safety. However, these systems still require large domain-specific datasets, which are scarce in psychotherapy due to privacy concerns and data sensitivity.

The rise of large language models (LLMs) [153] opens new opportunities for personalized, empathetic, and engaging digital psychotherapy [154] and mental health treatments [155]. Psychotherapy requires a real-time adaptation to subtle motivational, emotional, and behavioral cues. Unlike simpler NLP models, advanced LLMs like ChatGPT [65] are capable of generating responses that are naturally fluent, contextually appropriate, and emotionally attuned. Psychotherapeutic strategies such as reflective listening, a core element in MI and CBT, demand nuanced interpretation and reflective response, which requires far more than surface-level text manipulation such as paraphrasing or summarization. Furthermore, managing therapeutic dialogues involves pacing, topic transitions, goal alignment, and not merely following scripted dialogue flows.

Yet, applying LLMs in the psychotherapeutic context to develop LLM-powered or hybrid [151] psychotherapy chatbots still poses **three technical challenges**: **(1)** LLMs lack domain-specific knowledge to initiate and sustain conversations on specific *psychotherapeutic topics* (see Fig 4.3), such as increasing intrinsic motivation, or cognitive behavioral practices [20] like “mindfulness” [156], and cognitive distortions such as “should statements” [157]. **(2)** LLMs also struggle to generate structured and goal-directed *therapeutic questions* (see Fig 4.1) that guide behavioral interventions. Moreover, the reliance on fully expert-crafted dialogue scripts presents another challenge: **(3)** Creating these expert-crafted dialogue scripts is time-consuming and labor-intensive, posing barriers to scalability and broader adoption in real-world therapeutic contexts.

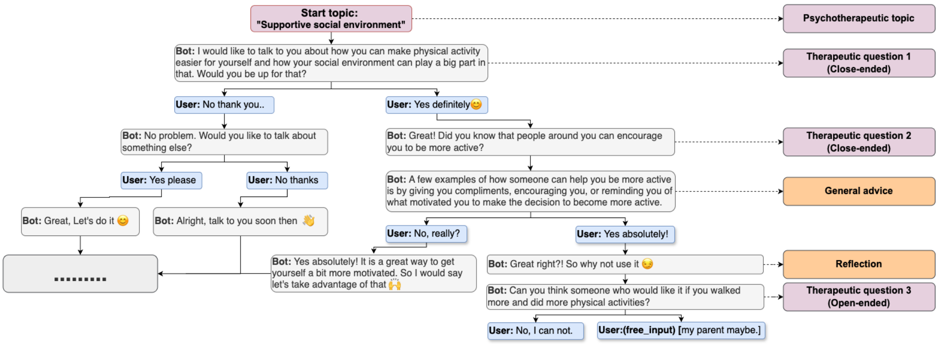


Figure 4.1: A dialogue example: parts of tree-structured dialogue scripts pre-crafted by experts under the psychotherapeutic topic "Supportive social environment" for behavioral intervention.

Despite the advancements of LLMs, it remains unclear whether expert-crafted dialogue scripts are still necessary to guide LLMs for psychotherapy, specifically in balancing generative flexibility with the structured and goal-oriented nature of evidence-based therapeutic conversations. This leads to our first research question. ***RQ1: Do expert-crafted dialogue scripts remain essential for chatbot-delivered psychotherapy in the era of LLMs?*** To explore this question, we propose Script-Aligned Generation (SAG), a concept that aligns LLMs with expert-crafted dialogue scripts to retain therapeutic structure while enabling conversational flexibility. To rigorously examine RQ1 and validate SAG, we define two sub-questions: ***RQ1.1: How do LLM-powered chatbots aligned with expert-crafted dialogue scripts compare to rule-based chatbots and pure LLMs for psychotherapy?*** and ***RQ1.2: How do LLMs aligned via fine-tuning and prompting differ in performing psychotherapy?*** To answer these questions and address the first two challenges, we conducted Study 1 (Fig 4.2.a), comparing four chatbot types: a rule-based chatbot using static expert-crafted scripts; a pure LLM without alignment; and two LLM-powered chatbots aligned with expert-crafted scripts (i.e., SAG) via either fine-tuning or prompting. Results showed that LLM-powered chatbots by SAG significantly outperformed both rule-based and pure LLMs across key assessing metrics, including linguistic quality, therapeutic relevance, empathy, engagement, MI adherence, and motivation enhancement. These findings highlighted the continuing importance of expert-crafted scripts in enabling LLMs to deliver engaging, safe, and therapeutically effective psychotherapy.

Although SAG aims to enable LLMs to balance conversational flexibility with therapeutic effectiveness by aligning them with expert-crafted dialogue scripts, it still relies on fully expert-crafted scripts, posing scalability and cost challenges for real-world psychotherapy applications. To address challenge (3), we proposed Script-Strategy Aligned Generation (SSAG), a more flexible and collaborative alignment approach. Unlike SAG, SSAG requires only partial expert input: core psychotherapeutic topics, key therapeutic questions, and optional advice (see Fig 4.8). Drawing on prior work in CSCW [158, 159] and NLP [160] that employ therapeutic strategies to guide human therapists and LLMs for facilitating effective therapy, SSAG enabled LLMs to dynamically generate dialogues aligned with these therapeutic strategies, such as asking questions, reflective listening, and giving advice, enabling alignments with both partial expert-crafted scripts and evidence-based therapeutic strategies [11, 159]. We therefore ask our second research question: ***RQ2: Can psychotherapy chatbots using SSAG achieve comparable conversational quality and therapeutic effectiveness to those using SAG?*** To evaluate this, we conducted Study 2 (Fig 4.2.b), a 10-day field study comparing three chatbots: (1) a rule-based chatbot as baseline, (2) an SAG-aligned chatbot (via prompting) as in Study 1, and (3) an SSAG-aligned chatbot. Results showed that SSAG matched SAG in therapeutic effectiveness while sustaining empathy and engagement, suggesting that flexible alignment can deliver comparable outcomes with less expert scripts.

To distinguish between SAG and SSAG, we defined SAG as strict alignment that relies on fully scripted dialogue content and flows, whereas SSAG is a flexible alignment approach that requires only partial expert input. SSAG enabled domain experts to contribute at a higher level of abstraction by specifying key therapeutic elements (e.g., psychotherapeutic topics and questions as illustrated in Fig 4.8) rather than scripting the entire dialogues. This leads to our third research question: ***RQ3: To what extent can SSAG reduce reliance on expert-scripted content for developing psychotherapy chatbots?*** Findings from Study 2 showed that SSAG chatbots achieve comparable performance to those using SAG while reducing the need for fully expert-crafted dialogue scripts, positioning LLMs as co-authors of therapeutic dialogues. SSAG is situated within CSCW by facilitating collaboration among psychotherapy experts, developers, and LLMs. This helps enable the scalable, co-designed development of chatbots that have the potential to be both engaging and therapeutically effective for health behavioral interventions.

We chose to focus on MI and CBT due to their strong evidence base and complementary structures. MI, a flexible, client-centered approach, emphasizes reflective listening and open-ended questioning to enhance intrinsic motivation [8, 9, 18, 19], making it well-suited for evaluating the ability of LLMs to support engaging and context-aware conversations. CBT, by comparison, offers structured, goal-oriented

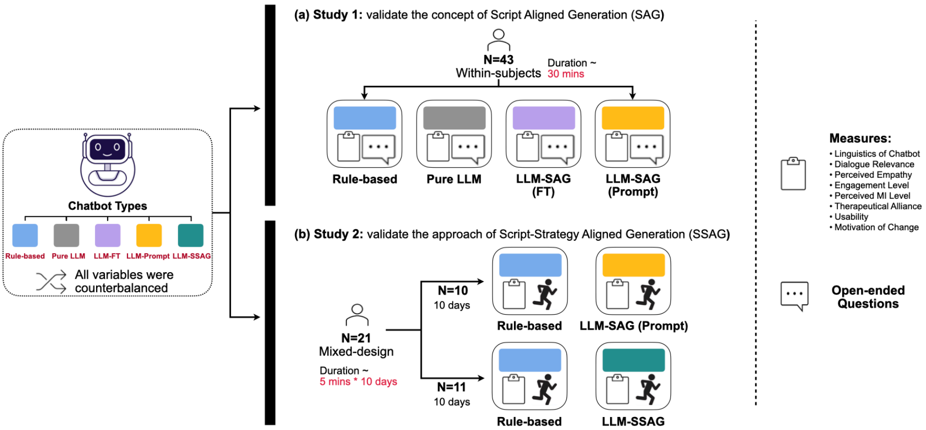


Figure 4.2: The procedure and design of the two studies conducted for evaluating the chatbots in delivering psychotherapy for health behavioral intervention.

dialogues like cognitive reframing [20], aligning well with LLMs guided by specific therapeutic objectives. By combining MI's flexibility with CBT's structure, we can better explore whether aligned LLMs can uphold therapeutic adherence without sacrificing perceived conversational quality such as empathy, a tradeoff investigated in prior work about CBT [21, 22]. Importantly, SSAG can be generalized beyond MI and CBT. By grounding generation in high-level therapeutic strategies, such as asking, reflecting, and advising, identified in prior CSCW work [158,159], SSAG can be extended to other frameworks like Acceptance and Commitment Therapy (ACT) [161] or digital mental health coaching [162, 163] by substituting context-appropriate, expert-defined strategies. This positions SSAG as a scalable, adaptable approach for a wide range of evidence-based digital health interventions.

The work in this chapter contributes to CSCW by advancing the effective development of LLM-powered psychotherapy chatbots for health interventions. First, we proposed and evaluated Script-Aligned Generation (SAG), a concept that aligns LLMs with fully expert-crafted scripts to balance therapeutic effectiveness and conversational flexibility, highlighting the ongoing importance of human expertise in guiding LLMs. Second, we introduced Script-Strategy Aligned Generation (SSAG), a more flexible approach that reduces reliance on expert-scripted content by requiring only partial human input (e.g., therapeutic topics and questions), enabling efficient development of psychotherapy chatbots supported by LLMs. Third, we contributed the first dataset of expert-authored psychotherapy dialogues grounded in evidence-based techniques (MI and CBT), supporting future research on LLM alignment for psychotherapy. Together, these contributions demonstrate that expert guidance remains critical for safe and effective digital psychotherapy, and that SSAG offers a practical path toward scalable, controllable, and efficient development of therapeutically aligned chatbots in the era of LLMs.

4.2 Related Work

4.2.1 Chatbots and Conversational Design for Digital Psychotherapy and Behavioral Intervention

The application of chatbots or conversational agents (CAs) in delivering psychotherapy and health- interventions has gained significant attention as a means of broadening access to psychotherapeutic and mental health support services like health behavioral intervention [4,6,147] and mental healthcare [148,149]. Rule-based chatbots [12], which operate on predefined dialogue scripts, are widely used in digital psychotherapy for their high controllability and explainability, essential qualities in sensitive fields such as mental healthcare and psychotherapy. These systems ensure precision in addressing therapeutic needs and prevent deviations from clinically validated pathways. For example, studies have used rule-based chatbots for Motivational Interviewing (MI) [8, 9] and Cognitive Behavioral Therapy (CBT) [20] to support interventions like smoking cessation [4, 6] and physical activity promotion [147, 164], where controlled, structured responses are crucial.

Despite the benefits of controllability and transparency, rule-based chatbots face limitations in dialogue empathy and flexibility. Their interactions tend to be rigid, lacking the dynamic adaptability of human therapists, which is particularly essential in psychotherapy, where a nuanced understanding of user queries and empathetic responses is critical [158]. For instance, a core principle of MI or CBT is reflective listening [11], where the therapist mirrors the client's expressions to promote self-reflection and insight. However, rule-based chatbots struggle to emulate such responsive empathy [7], as their responses are confined to predefined scripts that may not fully capture the depth of user needs or emotional nuances. More importantly, rule-based chatbots typically rely on intent-based dialogue systems [12], where domain experts design expert-crafted scripts to define chatbot's response to various user intents, ensuring consistent therapeutic guidance and simulating therapeutic conversations with a high degree of accuracy and reliability. However, this dependence on expert-written scripts makes rule-based chatbots highly resource-intensive, as they require extensive input from experts to develop domain-specific dialogue scripts [13, 14, 165]. Implementing these expert-crafted dialogues often involves encoding intricate conversation designs, including conversational content and flows, that incorporate strategies from psychotherapy such as MI or CBT, making the development process time-consuming and costly. Therefore, our work created a dataset with extensive expert-crafted dialogue scripts tailored for physical activity interventions using MI and CBT.

4.2.2 Language Models for Digital Psychotherapy and Behavioral Intervention

The integration of generative language models [91, 154] into digital psychotherapy shows promise in enhancing conversational flexibility and depth in chatbot-driven therapeutic interventions. Initial efforts to improve rigid, rule-based chatbots led to hybrid systems from prior work by [151] that combined Natural Language Generation (NLG) [150] with predefined dialogue scripts. These systems aimed to improve conversational fluidity by incorporating generative elements to create more empathetic, reflective responses, a core principle of many psychotherapeutic techniques such as MI [8, 9] and CBT [20]. For example, hybrid models generate reflective statements or paraphrases to encourage clients to explore their thoughts while staying aligned with therapeutic goals [90, 151, 152]. With advancements in LLMs [153, 166], the potential for more adaptive and contextually aware dialogues in digital psychotherapy has grown considerably. LLMs offer a powerful ability to generate varied and nuanced responses, which can help maintain engagement and improve the user experience during psychotherapy [160, 167]. In MI or CBT, where client engagement and reflective listening are critical [7], LLMs have shown promise by producing responses that align with therapeutic techniques [21, 160], with a more flexible and engaging conversational style [155, 158, 168]. A recent study [155] demonstrated the effectiveness of generative AI chatbots in providing clinical-level mental health treatments, supported by evidence from a long-term randomized controlled trial.

Despite these advantages, the use of LLMs in psychotherapy presents considerable challenges [169], especially concerning technical controllability, model transparency [170], cultural understanding [21] and ethical concerns [37] in sensitive contexts. Unlike rule-based systems that follow experts' pre-defined guidance or expertise, LLMs operate with a black-box nature that generates responses based on probabilistic models [171] that may lack explicit adherence to therapeutic principles or ethical standards [37]. This lack of predictability and controllability poses risks, as LLMs may inadvertently generate responses that are insensitive, ethically inappropriate, or even harmful, especially in sensitive interactions where mental health and emotional well-being are involved. Furthermore, LLMs struggle to consistently apply psychotherapeutic counseling techniques, such as MI or CBT, as they lack the ability to systematically follow therapeutic strategies without expertise guidance. These challenges have prompted further investigations into how LLMs and generative AI systems can be instructed or aligned with therapeutic principles [158, 160, 172], human values [173–175], external knowledge [176], and ethical standards [177, 178] to ensure safe and effective deployment in sensitive contexts such as psychotherapy for health intervention.

4.2.3 Aligning LLMs with Domain Expertise and Instructions for Psychotherapy

Recent advances in aligning LLMs has evolved beyond linguistic fluency to emphasize adherence to goal-directed and domain-informed dialogue. This shift from purely probabilistic generative models [171] to more adaptive systems has introduced approaches such as instructed dialogue generation [179], reinforcement learning from human feedback (RLHF) [174, 175], and context-aware generation [180]. These approaches enable LLMs to tailor response generation based on specific conversational strategies, user intent, and domain objectives [160, 172, 181]. This also enables more effective mixed-initiative dialogue [176, 182], where both the user and model collaboratively guide the conversational interaction.

In psychotherapy, LLMs offer transformative potential to deliver flexible and engaging interventions; however, prior work by Iftikhar et al. [21] indicates that instructed LLMs may overly rely on therapeutic techniques with compromising conversational qualities such as empathy. To address this, alignment approaches should guide LLMs to balance conversational quality with adherence to structured therapeutic principles, which makes the alignment of LLMs with domain expertise critical to ensure safety, empathy, engaging, and goal-directed conversations. Aligned dialogue generation is expected to support models to embed psychological and empathetic principles within responses [155, 183–185], allowing LLMs to effectively emulate the precision of rule-based systems while retaining the flexibility needed for nuanced, context-sensitive support in psychotherapeutic settings.

Two dominant alignment approaches have emerged: fine-tuning and prompting. Fine-tuning [186] integrates domain expertise directly into model weights, offering strong domain fidelity [21, 155, 172]. However, it is resource-intensive and dependent on large, sensitive datasets that are difficult to obtain in psychotherapy. Prompting [187], especially with in-context learning methods like Chain-of-Thought [188–190] and Tree-of-Thoughts [191], provides a more scalable and flexible alternative, guiding LLM outputs with structured cues instead of continuous model retraining. Recent work [160, 192] has shown that incorporating therapeutic strategies into prompts allows LLMs to produce more therapeutic-aligned responses [158, 159]. Comparative studies [193] also validate both prompting and fine-tuning as effective approaches for mental health tasks, supporting their potential in psychotherapeutic interventions.

Despite this progress, these approaches typically assume access to comprehensive domain-specific data for aligning the LLMs, which can be costly and time-consuming for domain experts to produce. To address this, we proposed Script-Strategy Aligned Generation (SSAG) in this work, an alignment approach that requires only partial expert-crafted dialogue scripts, thereby reducing reliance on fully scripted dialogue content or flows, while using LLMs to dynamically manage dialogue flow and generate responses. SSAG supports a collaborative expert-LLM co-authoring workflows and enables a controllable alignment through the step-wise therapeutic strategy prediction mechanism. Compared to prompting and fine-tuning, SSAG offers a more flexible and efficient alternative for the real-world development of psychotherapy chatbots, especially in the contexts where data is limited and expert labor is costly.

4.3 Creating Dataset with Expert-Crafted Dialogue Scripts

A team of fifteen experts (i.e., research associates, each holding a BSc or higher degree in clinical or health psychology), collaboratively developed a comprehensive dataset of pre-scripted, tree-structured dialogues designed to support physical activity interventions. This dataset was specifically designed to serve as input for developing conversational agents in this study and the broader TIMELY project, ensuring both therapeutic relevance and practical applicability in health behavioral interventions. These research associates were involved in the Timely study and were responsible for supervising research interns during the dataset creation process. Their academic background and hands-on involvement in behavioral intervention research provided them with the necessary foundation to design realistic and therapeutically informed dialogue flows. Pre-scripted dialogues include both dialogue content and flows in tree-structure as shown in Fig 4.1. Grounded in the psychotherapeutic techniques of MI [8, 9] and CBT [20], these scripts target common therapeutic scenarios in behavioral interventions. Examples of crafted dialogues are shown in Fig 4.3, covering topics such as “rating confidence of change” in MI and “supportive social environment” in CBT.

The creation process began with brainstorming sessions, where experts identified psychotherapeutic topics relevant to behavioral interventions, focusing on physical activity. Topics like overcoming exercise barriers and improving sleep were chosen to cover diverse intervention scenarios. Each expert then crafted dialogues for specific topics, ensuring the dataset captured the depth and variability of real therapy while adhering to MI and CBT principles. To ensure clarity and authenticity, a think-aloud protocol [194] was used.

The final dataset includes over 1,800 expert-written utterances across 26 **psy-**

Psychotherapy Technique	Psychotherapeutic Topic	Example of Expert-crafted Dialogues	Therapeutic Strategy
Motivational Interviewing (MI)	"Rating Confidence of Change"	**Therapist**: So my first question is: on a scale from 0 to 10, how confident are you that you will reach your step goals tomorrow 🎯?	Therapeutic Question 1
		Client: I would rate as 7 maybe.	
		Therapist: Thanks for sharing with me😊 Seems like you're already quite confident about reaching your goal. Can you tell me why you gave it a 7 but not a 5? Any aspects make you confident right now to achieve your step count goal?	Reflection Therapeutic Question 2
		Client: Because I have to walk to school every day, so I should be able to have enough steps.	
		Therapist: That's indeed something that could make you more confident! Do you know what would need to change so you will rate your confidence a 9?	Reflection Therapeutic Question 3
		Client: Hmm, I don't know …	
		Therapist: I understand.. trying to raise your confidence is easier said than done. But I am here to help! I have a few small tips you can try to boost your confidence to reach your step count goal. Would you like to know?	Reflection Therapeutic Question 4
		Client: Sure, why not.	
		Therapist: The best way to gain more confidence is to see yourself succeed, so the first tip is: split your goal into small pieces and try to achieve those! My second tip is: ……	General Advice
Cognitive Behavioral Therapy (CBT)	"Supportive Social Environment"	**Therapist**: I would like to talk to you about how your social environment can play a big part in making your physical active easier. Did you know that people around you can encourage you to be more active?	Therapeutic Question 1
		Client: Maybe there are someone …	
		Therapist: Great! So why not use it😏 Can you think someone who would like it if you walked more and did more physical activities?	Reflection Therapeutic Question 2
		Client: I think, my mum would like that if I can walk more or do more physical activities.	
		Therapist: Nice! This person can be a great supporter in your journey to become more active 👫 Can you think of two ways in which he or she can be helpful?	Reflection Therapeutic Question 3
		Client: Hmm, can you maybe have an example?	
		Therapist: Of course! An example of how someone can help you be more active is by giving you compliments, encouragement, or reminding you of what motivated you to make decision to be more active.	General Advice

Figure 4.3: Examples of expert-crafted dialogues in motivational interviewing (MI) and cognitive behavioral therapy (CBT) with different psychotherapeutic topics. The expert-crafted dialogues contain 1) psychotherapeutic topics, 2) therapeutic questions, 3) reflections, 4) general advice, and 5) dialogue flows pre-designed as in Fig 4.1.

chotherapeutic topics as detailed in 4.A. The dataset includes a mix of closed- and open-ended questions, reflections, and general advice, key elements for eliciting client self-reflection and promoting engagement, which are core objectives of both MI and CBT. Each topic is structured by a series of **therapeutic questions**. The tree-structured format as shown in Fig 4.1 ensures a coherent flow and branching, designed to mirror real-world psychotherapeutic dynamics. More detailed dialogue examples are provided in Appendix 4.B and 4.C

4.4 Study 1: Concept of Aligning LLM with Full Expert-Crafted Dialogue Scripts

4.4.1 Concept Validation: Script-Aligned Generation (SAG)

Study 1 aimed to validate the concept of Script-Aligned Generation (SAG), which aligns LLMs with full expert-crafted dialogue scripts for psychotherapy. SAG seeks to balance therapeutic fidelity with the conversational flexibility and engagement inherent in LLMs. The goal of Study 1 was to assess SAG's potential to enhance LLMs' potential to deliver flexible, engaging, and expert-guided psychotherapy. To implement the concept of SAG, we proposed two alignment approaches: fine-tuning (LLM-SAG (FT)) [186] and prompting (LLM-SAG (Prompt)) [187], to align LLMs with expert-crafted, tree-structured dialogue scripts designed for health intervention as shown in Fig 4.1, aiming to determine which provides the best combination of therapeutic effectiveness and conversational qualities.

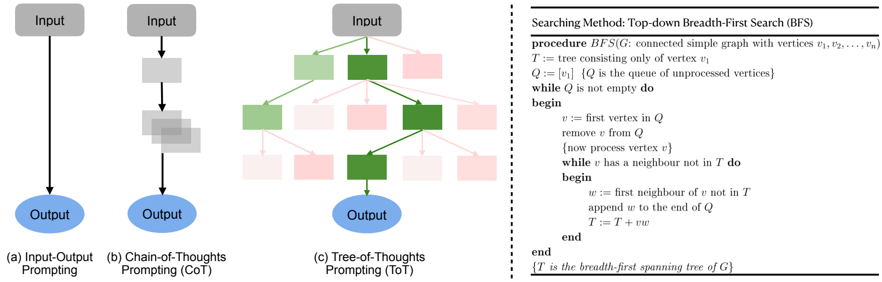


Figure 4.4: Visualization of aligning LLMs with expert-crafted tree-structured dialogue scripts via Tree-of-Thoughts prompting [191].

4.4.2 Study Methods

Study design and procedure

A comprehensive overview of the study procedure is illustrated in Fig 4.2 (a). We conducted an evaluation study employing within-subjects design to compare four chatbot types: 1) a rule-based chatbot ("rule-based") strictly followed the expert-crafted dialogues; 2) a pure LLM ("pure LLM") without any alignment with expert-crafted dialogues; and LLM-powered chatbots strictly aligned with the full expert-crafted scripts through either 3) prompting ("LLM-SAG (Prompt)") or 4) fine-tuning ("LLM-SAG (FT)").

Before the study, each participant received an information letter and provided informed consent. After, participants interacted with each chatbot in counterbalanced orders and completed a survey immediately after interaction to assess their experience with that specific chatbot. This within-subjects design allowed for a direct comparison of user assessments across chatbot types under consistent conditions, aiming to identify which type of chatbot best performs in facilitating the therapeutic interactions.

Types of chatbots

Study 1 examined the following four types of chatbot:

1) **rule-based chatbot.** This chatbot was implemented using the RASA framework [195], a popular open-source platform for building rule-based chatbots. It strictly followed the expert-crafted dialogue scripts detailed in Section 4.3, using predefined intents, responses, and dialogue flow rules encoded in RASA's dialogue management model [195]. This chatbot did not generate responses beyond what is explicitly pre-scripted, ensuring high controllability. This condition served as a

baseline, representing a traditional chatbot approach where all conversational paths are predetermined by experts.

2) **pure LLM.** This chatbot used a GPT-4o model [196] without alignment to expert-crafted dialogue scripts. At the start of each session, the LLM was prompted with: "You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy). The current therapeutic topic is [given topic]. The definition of this topic is [description of the given topic]." The topic was selected from a set of expert-defined **psychotherapeutic topics** (as demonstrated in Section 4.3 and detailed in 4.A). Participants can interact with this chatbot freely within the specified topic. This setup evaluated the pure LLM's performance in delivering psychotherapy with high-level instructions, but without script guidance or alignment.

3) **LLM-SAG (FT).** We fine-tuned a GPT-4o model using our expert-crafted dialogue dataset, which consisted of tree-structured dialogues authored and reviewed by health psychology experts for clarity and coherence (see Section 4.3). We employed supervised fine-tuning involving input-output pairs, ensuring the LLM's responses aligned with expert-crafted dialogue responses and flows. Additionally, to evaluate the chatbot's adherence to the predefined dialogue flow, participants can ask "out-of-script" questions to test the chatbot's ability to respond appropriately while maintaining alignment to the specific dialogue flow. This method also applied to the LLM-SAG (Prompt) chatbot.

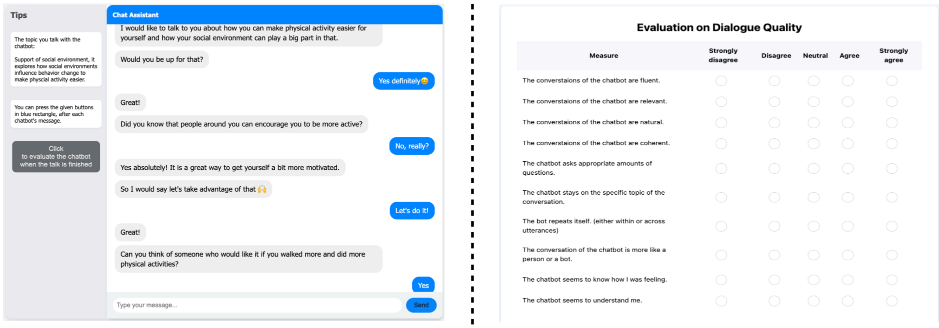
4) **LLM-SAG (Prompt).** For a more scalable and computationally efficient alternative to fine-tuning [197], we implemented a prompting-based alignment using the Tree-of-Thoughts (ToT) technique [191], as visualized in Fig 4.4. ToT guided the GPT-4o model through our pre-authored, tree-structured dialogue scripts using a Breadth-First Search (BFS) algorithm [198]. We chose ToT for its effectiveness to align LLM with tree-structured expert-crafted dialogue scripts. Unlike other prompting techniques (e.g., few-shot or chain-of-thought [188, 189]), ToT ensures consistency in multi-turn dialogue scenarios by enabling controlled, step-by-step navigation through predefined tree-based dialogue branches. At each turn, the LLM was prompted with the current dialogue context and a constrained set of valid next-turn response options through predefined dialogue branches in the tree structure, preserving the dialogue logic and therapeutic progression embedded in expert scripts. Compared to fine-tuning, ToT is a more lightweight and flexible approach that avoids retraining while maintaining integrity to expert scripts. While ToT was well-suited for aligning with tree-structured scripts in this work, we did not compare it with alternative prompting strategies (e.g., chain-of-thought). Future work could

evaluate how different prompting strategies perform in this context.

The prompting templates used for the LLM-powered chatbots are provided in Appendix B. The implementation details of the models and settings are presented in 4.G.

Web-based interfaces for chatbot interaction and evaluation

The study was conducted using a self-developed web-based application that integrated both chatbot interaction and an evaluation survey into one seamless interface, as shown in Fig 4.5. This design allowed participants to interact with the chatbot and provided immediate feedback in a continuous session, streamlining user experience. The interface was optimized for simplicity, minimizing distractions and ensuring smooth navigation, enabling participants to focus on their interactions without difficulties.



Chat Interface **Evaluation Interface**

Figure 4.5: Web-based interfaces were developed for Study 1 to enable chatbot interactions and facilitate immediate evaluation.

Measures

Participants' baseline exposure to chatbots was assessed using four 5-point Likert-scale items: (1) frequency of chatbot use (1 = Never, 5 = Daily), (2) prior experience with chatbots (1 = Strongly negative, 5 = Strongly positive), (3) familiarity with chatbots (1 = Not familiar at all, 5 = Extremely familiar), and (4) general attitude toward chatbots (1 = Strongly negative, 5 = Strongly positive).

In addition, we assessed above four chatbot types using the measures as follows. All quantitative self-report measures used a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

Linguistic Quality: Evaluated the fluency, relevance, naturalness, coherence, and the human-likeness of the chatbot's responses, using five assessing items adopted from [199, 200]. Example items are "The conversations of the chatbot are fluent." and "The conversation of the chatbot is more like a person rather than a bot."

Dialogue Relevance: Inspired by work [201], two self-constructed items were used to specifically assess the alignment of expert-crafted therapeutic dialogues. Example items are “The chatbot asks an appropriate amount of questions.” and “The chatbot stays on the specific topic of the conversation.”

Empathy and Engagement: Gauged perceived empathy of the chatbot’s conversations, which is crucial for therapeutic effectiveness. We adopted the questionnaire from [7,202] with two items: “The chatbot seems to know how I felt.” and “The chatbot seems to understand me.” To assess the chatbot’s ability to maintain user interest and interaction as the engagement level, a questionnaire from [203] was adopted. Example items are “I lost myself in the interaction with this chatbot.” and “The chatbot is enjoyable to talk to.”

Perceived Motivational Interviewing (MI) Adherence: Evaluated how well the chatbot simulates an MI session and adheres to the MI principles. We adopted a questionnaire from [7, 204] with four items. Example items are “The chatbot helped me talk about changing my behavior.” and “The chatbot helped me feel hopeful about changing my behavior.”

Motivation for Change: Measured the chatbot’s impact on participants’ motivation to change, a key therapeutic outcome. We self-constructed an item “I am motivated to make changes in the behavior after I talk with the chatbot.”

Therapeutic Alliance: Explored the rapport and connection between the chatbot and the user. We adopted the questionnaire from [205] with six items. Example items are “I believe this chatbot can help me to address my problem.” and “This chatbot encourages me to accomplish tasks and make progress for the change we discussed.”

Usability: Measured the ease of interaction with the chatbots. We adopted the questionnaire called Bot Usability Scale from [206] with seven items. Example items are “Communicating with the chatbot was clear.” and “The chatbot’s responses were easy to understand.”

Two Open-Ended Questions were included to capture participants’ feelings and opinions about chatbots: “What have you enjoyed most or least about interacting with this chatbot?” and “What could be improved about this chatbot?”

Automatic Evaluation Metrics: We also developed two self-defined, tailored metrics to evaluate the chatbots’ performance in delivering expert-crafted *psychotherapeutic topics* and *therapeutic questions* (shown in Fig 4.1 and Fig 4.3). Both metrics were reported as relative proportions (0–100%) to allow for consistent comparison across different chatbots. *“Auto-Metric 1”* assesses the number of psychotherapeutic topics that reached a natural conclusion, reflecting the level of user engagement with chatbots. *“Auto-Metric 2”* measures the total number of expert-crafted therapeutic questions asked by the chatbot, indicating its effectiveness in guiding the conversation toward specified therapeutic goals.

Participants

Demographic	Categories	Numbers of Participants (%)
Gender	Female	22 (51.2%)
	Male	21 (48.8%)
Age	18-24	23 (53.5%)
	25-34	14 (32.6%)
	35-44	2 (4.7%)
	45-54	3 (7.0%)
	65+	1 (2.3%)
Education	High school degree or equivalent	8 (18.6%)
	Bachelor's degree	15 (34.9%)
	Master's degree	16 (37.2%)
	Doctorate or higher	4 (9.3%)
Professional Domain	Health and Medical Science	7 (16.3%)
	Science, Technology, Engineering, Mathematics (STEM)	14 (32.6%)
	Business, Economics, and Law	6 (14.0%)
	Communication, Arts, Culture and Entertainment	5 (11.6%)
	Education and Social Science	8 (18.6%)
	Government and Public Sector	2 (4.7%)
	Other	1 (2.3%)

Table 4.1: Characteristics of participants in Study 1.

4

Participant demographics are shown in Table 4.1. A power analysis using G*Power [207] indicated that at least 30 participants were required to detect a medium effect size (d = 0.25) and $\alpha = 0.05$ with 90% power. To ensure robustness, we recruited 43 participants through institutional channels and social media. Participants had a diverse demographic profile. Eligibility criteria required participants to be at least 18 years old and fluent in English. Participation was voluntary, and participants were compensated with money or study credits for completing a 30-minute online session. This study was approved by the institutional ethics committee at the University of Amsterdam.

Data analysis

To analyze the impact of different chatbots on participants' perceptions of psychotherapeutic interventions, we first tested the data's suitability for statistical analysis. Normality was assessed using the Shapiro-Wilk test [208], and homogeneity of variance via Bartlett's test [209]. As the data violated normality assumptions, we applied a Generalized Estimating Equation (GEE) model [210] to compare mean scores across chatbot conditions. Post-hoc pairwise comparisons between the LLM-SAG (Prompt) and LLM-SAG (FT) chatbots were conducted using Wilcoxon signed-rank tests [211] with Bonferroni correction [212]. Additionally, Spearman correlation analysis [213] was used to examine whether participants' prior exposure to chatbots was associated with key outcome measures.

For the qualitative data, we conducted an inductive content analysis [214] of responses to two open-ended questions. The first two authors developed an initial codebook using ATLAS.ti [215] based on participants' perceptions and suggested improvements for each chatbot type. Both coders independently coded the responses, revising the codebook as new themes emerged. Codes were refined, merged where appropriate, and re-coded to ensure consistency.

4.4.3 Quantitative Findings

Automatic evaluation metrics

Chatbot Types	Auto-Metric 1	Auto-Metric 2
Rule-based (as oracle)	100.00	98.81
Pure LLM	85.71	12.62
LLM-SAG (FT)	71.43	70.24
LLM-SAG (Prompt)	97.62	96.42

Table 4.2: Results (ratio) of automatic evaluation metrics for rule-based, pure LLM, and LLM-SAG chatbots.

Table 4.2 shows the results of automatic evaluation metrics, as demonstrated in Section 4.4.2. For "Auto-Metric 1" (i.e., measuring completion of each psychotherapeutic topic), rule-based chatbot scores the highest, followed by LLM-SAG (Prompt) and pure LLM, which have similar scores. In "Auto-Metric 2" (i.e., measuring therapeutic questions asked), the rule-based chatbot again leads, with LLM-SAG (Prompt) close behind. The pure LLM scores the lowest, indicating a significant deviation from the expert-crafted dialogue scripts.

Descriptive statistics

Participants reported moderate prior exposure to chatbots. The average frequency of chatbot use was 3.14 (SD=0.83), prior experience was 3.23 (SD=1.08), familiarity with chatbots was 3.44 (SD=0.73), and overall attitude toward chatbots was 3.40 (SD=0.79). Across the four chatbot types: rule-based, pure LLM, LLM-SAG (FT), and LLM-SAG (Prompt), the LLM-SAG (Prompt) consistently outperformed others across multiple measures.

As shown in Table 4.3, LLM-SAG (Prompt) achieved the highest ratings in linguistic quality (M=3.95, SD=0.23), dialogue relevance (M=3.39, SD=0.68), and empathy (M=3.07, SD=0.16), and showed strong performance in engagement (M=2.90, SD=0.33), perceived MI adherence (M=3.74, SD=0.13), motivation (M=3.42, SD=0.00), usability (M=3.82, SD=0.26), and therapeutic alliance (M=3.58, SD=0.46), surpassing other chatbot types. This indicated that it not only followed expert-crafted dialogue scripts closely but also delivered a more flexible and effective therapeutic experience overall.

Comparison	Measure	Mean (SD)	Coefficient	Effect ($Std.B$)	p-value
Rule-based vs. Pure LLM					
	Linguistic Quality	3.66 (.22) vs. 3.76 (.14)	.10	.14 (small)	.45
	Dialogue Relevance	3.26 (.88) vs. 3.09 (.33)	-.13	.22 (medium)	.34
	Empathy	2.53 (.10) vs. 2.99 (.18)	.45	.44 (medium)	.04 *
	Engagement	2.58 (.53) vs. 2.72 (.42)	.14	.18 (small)	.33
	Perceived MI	3.13 (.15) vs. 3.13 (.08)	.00	.00 (small)	1.00
	Motivation Change	2.86 (.00) vs. 3.00 (.00)	.14	.13 (small)	.47
	Therapeutic Alliance	3.31 (.64) vs. 3.33 (.36)	.02	.02 (small)	.92
	Usability	3.60 (.40) vs. 3.56 (.24)	-.04	.05 (small)	.82
Rule-based vs. LLM-SAG (FT)					
	Linguistic Quality	3.66 (.22) vs. 3.60 (.18)	-.06	.08 (small)	.65
	Dialogue Relevance	3.26 (.88) vs. 3.38 (.51)	.15	.25 (medium)	.21
	Empathy	2.53 (.10) vs. 2.94 (.15)	.41	.40 (medium)	.04 *
	Engagement	2.58 (.53) vs. 2.58 (.54)	.00	.00 (small)	1.00
	Perceived MI	3.13 (.15) vs. 3.31 (.12)	.19	.17 (small)	.41
	Motivation Change	2.86 (.00) vs. 3.00 (.00)	.14	.13 (small)	.46
	Therapeutic Alliance	3.31 (.64) vs. 3.33 (.48)	.01	.01 (small)	.94
	Usability	3.60 (.40) vs. 3.61 (.26)	.01	.01 (small)	.97
Rule-based vs. LLM-SAG (Prompt)					
	Linguistic Quality	3.66 (.22) vs. 3.95 (.23)	.30	.40 (medium)	.02 *
	Dialogue Relevance	3.26 (.88) vs. 3.39 (.68)	.16	.27 (medium)	.20
	Empathy	2.53 (.10) vs. 3.07 (.16)	.53	.52 (large)	.01 **
	Engagement	2.58 (.53) vs. 2.90 (.33)	.32	.40 (medium)	.03 *
	Perceived MI	3.13 (.15) vs. 3.74 (.13)	.62	.57 (large)	.01 **
	Motivation Change	2.86 (.00) vs. 3.42 (.00)	.56	.51 (large)	.01 **
	Therapeutic Alliance	3.31 (.64) vs. 3.58 (.46)	.26	.32 (medium)	.10
	Usability	3.60 (.40) vs. 3.82 (.26)	.22	.30 (medium)	.19

Table 4.3: Results of generalized estimating equations (GEE) [210] comparing rule-based with three types of LLM-powered chatbots (pure LLM, LLM-SAG via prompting and fine-tuning). (**p<.01, *p<.05, "Coefficient" represents the unstandardized regression coefficient.)

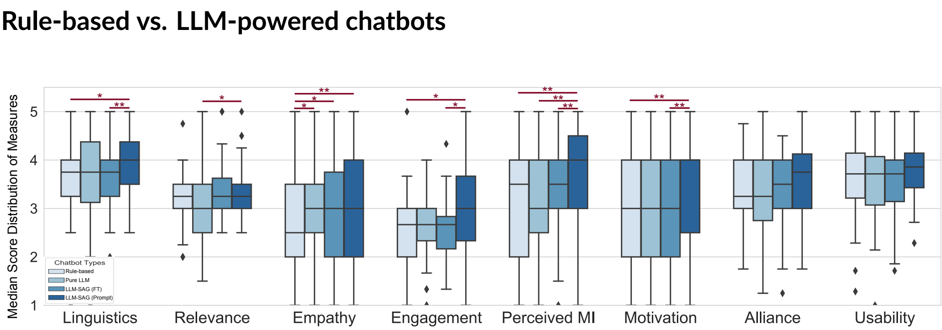


Figure 4.6: Distribution of median value of measures in Study 1 across four chatbot types. Y-axis ranges from 1 to 5 (Likert scale). Each box shows the interquartile range (25th–75th percentile), with the midline indicating the median. Whiskers extend to 1.5 times the interquartile range, and dots outside represent outliers. Horizontal red lines with asterisks indicate statistically significant differences between conditions (*p < .05, **p < .01).

To address RQ1.1, we compared the performance of rule-based and LLM-powered chatbots, including pure LLM, LLM-SAG (FT) and LLM-SAG (Prompt). As depicted in Table 4.3 and Fig 4.6, LLM-SAG (Prompt) significantly outperformed the rule-based chatbot in linguistic quality, empathy, engagement, perceived MI adherence, and motivation change. Pure LLM and LLM-SAG (FT) chatbots only showed significantly higher empathy ratings than the rule-based chatbot (p=0.04) but performed similarly or even worse in dialogue relevance and MI adherence. LLM-SAG (Prompt) chatbot scored significantly higher than pure LLM in dialogue relevance and perceived MI shown in Fig 4.6. Moreover, all chatbots performed similarly in usability and therapeutic alliance. These findings affirmed the value of aligning LLMs with expert-crafted scripts to preserve therapeutic quality and conversational engagement.

Prompting (LLM-SAG (Prompt)) vs. Fine-tuning (LLM-SAG (FT))

Comparison	Measure	Mean (SD)	Effect ($Std.B$)	p-value
LLM-SAG (Prompt) vs. LLM-SAG (FT)				
	Linguistic Quality	3.95 (.23) vs. 3.60 (.18)	.43 (large)	.01 $**$
	Dialogue Relevance	3.39 (.68) vs. 3.38 (.51)	.09 (small)	.54
	Empathy	3.07 (.16) vs. 2.94 (.15)	.07 (small)	.67
	Engagement	2.90 (.33) vs. 2.58 (.54)	.36 (medium)	.02 $*$
	Perceived MI	3.74 (.13) vs. 3.31 (.12)	.40 (medium)	.01 $**$
	Motivation Change	3.42 (.00) vs. 3.00 (.00)	.35 (medium)	.02 $*$
	Therapeutic Alliance	3.58 (.46) vs. 3.33 (.48)	.27 (medium)	.08
	Usability	3.82 (.26) vs. 3.61 (.26)	.24 (medium)	.11

Table 4.4: Results from Wilcoxon signed-rank test [211] for pairwisely comparing LLM-powered chatbots aligned with full expert-crafted dialogue scripts (SAG) through either prompting or fine-tuning. ($**p$<.01, $*p$<.05)

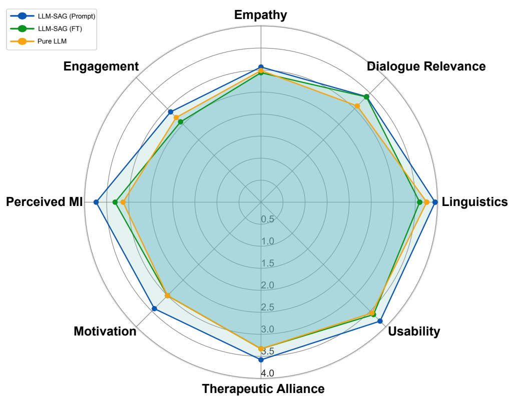


Figure 4.7: Visualization of the comparison (mean value) between the pure LLM chatbot and LLM-SAG chatbots through prompting (LLM-SAG (Prompt)) and fine-tuning (LLM-SAG (FT)). Each axis represents one of eight evaluation dimensions, with values plotted on a radial scale from 0 to 5 reflecting mean scores for each chatbot type.

To address RQ1.2, we compared two alignment approaches used to implement the concept of SAG: fine-tuning (LLM-SAG (FT)) and prompting (LLM-SAG (Prompt)). Results in Table 4.4 and Fig 4.7 showed that LLM-SAG (Prompt) significantly outperformed LLM-SAG (FT) in assessing diemensions, including linguistic quality (p=0.01), engagement (p=0.02), perceived MI (p=0.01), and motivation change (p=0.02). These findings suggested that prompting provides a more effective and flexible alignment approach for psychotherapy, likely due to its ability to flexibly navigate tree-structured dialogues while preserving expert-scripted dialogues for effective therapeutic experience, whereas fine-tuning may struggle to balance the adherence to the pre-crafted complex dialogue scripts and the adaptability to the "out-of-scripts" inputs from the participants.

4.4.4 Qualitative Findings: Open-ended Questions

A total of 344 free-text responses were collected in Study 1 from 43 participants across four chatbot types (i.e, rule-based vs. pure LLM vs. LLM-SAG (FT) vs LLM-SAG (Prompt)). 298 non-empty responses were included for analysis. Findings revealed four key themes shaping participant perceptions of chatbots in psychotherapy for health behavioral intervention that **specifically address RQ1**: (1) conversational quality and effectiveness, (2) balancing engagement with therapeutic adherence, (3) LLM-powered intelligences for digital psychotherapy, and (4) suggestions for improving chatbot-delivered psychotherapy.

Effective chatbot-delivered psychotherapy through clarity, relevance and coherence

In therapeutic contexts, the effectiveness of chatbot interactions depends heavily on two intertwined factors: clarity and coherence in conversations, and the therapeutic relevance of the content. This theme highlights the need for chatbots to respond in a way that is easy to understand, logically structured, and directly aligned with users' therapeutic goals.

The qualitative analysis showed that participants responded favorably to rule-based chatbots for their straightforward responses that were built on expert-crafted dialogue scripts. These chatbots were seen as effective in guiding users through therapeutic content with clarity. One participant remarked, "I liked the fact that the chatbot was really straightforward and clear in providing information, leading the users into questions. (rule-based chatbot)" However, participants also highlighted instances where clarity was lacking, especially in pure LLM-based chatbots. One participant shared, "It is not clear, and sometimes I don't know what I should say. Some examples are not clear at all. I wanted to see real instances, but it said ambiguously. (pure LLM)" This feedback highlights the need for clearer instructions, concrete examples, and more direct conversational guidance.

Therapeutic and contextual relevance were also viewed as essential. Some participants felt pure LLMs lacked focus, as one noted: "The topic is not quite relevant to the topic given; it would be good if it gave more specific advice on the topic we talked about for helping me in behavioral change." This highlights the importance of responses grounded in therapeutic objectives and user needs. While rule-based chatbots were appreciated for clarity, their rigidity was a frequent criticism. One participant noted, "I do not like chatbots with pre-set responses. There was no flexibility in what I could talk about." This inflexibility was seen as a barrier to deeper engagement. In contrast, LLM-based chatbots, particularly those aligned with expert-guided prompts (LLM-SAG), were praised for their adaptability and contextual responsiveness. As one participant described, "The chatbot can understand what I talk about and always stay within the topic given. It also asks relevant questions to move the talk forward. (LLM-SAG (Prompt))"

Maintaining logical flow and staying on topic was viewed as essential of effective interventions. Participants valued when chatbots responded coherently. One participant noted the improved coherence in LLM-SAG chatbots compared to rule-based one: "I can feel the dialogue from (rule-based) chatbot is similar to the previous one, and this time it (LLM-SAG (Prompt)) can understand what I said and give correct reply. I can feel it quite encouraging for change." Another participant emphasized the capability of aligned LLMs to stay on topic: "It guides my behavior change step by step in human-like conversations. And it is not overwhelming. It can understand my specific questions and give correct answers, and then it can go back to previous talk after my questions, interesting. (LLM-SAG (Prompt))"

Balancing conversational engagement and therapeutic adherence: Flexible vs. Structured

Conversational style plays a pivotal role in shaping users' perceptions and relationships with chatbots. Participants consistently highlighted how conversational style impacts their sense of engagement and therapeutic effectiveness.

The rule-based chatbot provided well-structured dialogues for psychotherapeutic intervention following expert-crafted content and was noted as, "I love it guiding my behavior change step by step. Instead of directly asking me to do certain things, it guide me to think about how to deliver such changes." However, participants described LLM-powered chatbots as more relatable and engaging than rule-based ones. The LLM-powered chatbots' ability to mimic natural conversation made interactions feel natural and fluid. As one participant noted, "The tone of this chatbot is interesting and appealing. It talks like a human being. (LLM-SAG (FT))" In contrast, rule-based chatbots with their rigid, pre-set responses were disliked in this context as, "I do not like (rule-based) chatbots with pre-set responses. There was no flexibility in what I could talk about with the chatbot." This lack of conversational flexibility left some participants feeling disconnected and constrained.

Interestingly, while pure LLM chatbots have been able to generate contextually relevant responses, participants noted that those aligned with expert-crafted dialogues provided more structured, professional, supportive and motivational conversations. These chatbots asked thoughtful questions, provided relevant examples, and offered advice in a structured, encouraging tone. One participant shared, "I liked the way it asked questions, gave examples and advices. I found that they were very relatable. (LLM-SAG (Prompt))" . This well-balanced flexibility and therapeutic structures for engaging and therapeutically-effective psychotherapy for behavioral interventions. However, some participants felt that overly professional or friendly tones could undermine the seriousness of certain topics. "I did not really like the super friend-like way of talking; I prefer a more professional type of way. (LLM-SAG (FT))" one participant noted. This points to the need for chatbots to dynamically adjust their conversational styles based on context, avoiding extremes and tailoring communication to user expectations and emotional needs.

In sum, the conversational styles of chatbots, particularly how they balance flexibility with structure, and friendliness with professionalism, directly influence user-perceived comfort, relatability, and engagement with the chatbots.

LLM-Powered Intelligence for Engaging Psychotherapy: Personalization, Memory, and Empowerment

This theme highlights the vital role of LLM-powered intelligence for enhancing digital psychotherapy for behavioral intervention, including features such as personalization, chatbot's memory, and user empowerment in creating engaging and effective chatbot-delivered interventions. The rule-based chatbot was disliked because of its pre-set responses and fixed dialogue structures without personalization, as noted, "I enjoy receiving the intended lesson and information in a very straightforward manner. [...] However, it did feel less personal and tailored. It was more like a good lesson than a conversation and coaching." Most participants appreciated chatbots that responded in a personalized but still structured way, offering tailored advice rather than generic responses. One participant remarked, "It guides my behavior change step by step in human-like conversations. And it is not overwhelming. It can understand my specific questions and give correct answers, and then it can go back to the previous talk after my questions, interesting. (LLM-SAG (Prompt))" This kind of personalized guidance, delivered at a comfortable pace, helped participants feel heard, supported, and empowered in their behavioral change journey.

In addition, participants also emphasized that chatbots capable of remembering prior interactions and adapting their responses accordingly made the conversations more coherent, supportive, and relatable. One participant shared, "The initial advice was also very generic and not useful for me specifically but in the rest of the conversation, the bot switched to suggestions, which were much more tailored to me and actually did inspire me. I even made a note and am going to implement one of the suggestions. (pure-LLM)" This illustrates how contextual awareness can strengthen engagement and empower the effectiveness of psychotherapeutic interventions. This ability to build on past interactions can also enhance the continuity and personalization of long-term conversations.

Therefore, these LLM-powered intelligent features could help build more engaging, empowering, and effective chatbot-delivered interventions. By integrating these features and maintaining the right balance of conversational styles, chatbots can better support users' therapeutic goals and foster more engaging experiences.

User-Informed Suggestions for Enhancing Chatbot-Delivered Psychotherapy

Participants shared valuable suggestions to improve chatbot-delivered psychotherapy for health behavior change, emphasizing the need for more engaging, adaptable, and therapeutically effective interactions.

A key suggestion was to integrate LLM capabilities into rule-based systems. Participants believed this would combine the structured guidance of scripted dialogues with the conversational flexibility, pointing to the potential for more dynamic and personalized interactions. Alongside this, participants highlighted the value of open-ended questions, which LLMs are well-suited to generate. Closed-question formats were seen as limiting, while open dialogue can encourages reflection and deeper engagement in the therapeutic process.

Participants also stressed the importance of modulating the chatbot's conversational style. A balanced tone, friendly yet professional, was viewed as critical. One participant shared, "It could have used a slightly different tone to balance the formality that enhances its reliability and friendliness."

Personalization emerged as another vital area for improvement. Participants wanted chatbots to tailor responses to their specific circumstances and offer actionable advice. One participant noted, "It should have given clearer instructions for users for their change" highlighting the need for more individualized support. In line with this, participants emphasized the importance of memory and contextual awareness. The ability of a chatbot to recall past interactions was seen as essential for maintaining continuity, deepening personalization, and improving therapeutic relevance.

Collectively, these user insights point toward a future of chatbot-delivered therapy that is more personalized, flexible, engaging, and therapeutically effective, ultimately leading to more impactful psychotherapeutic interventions.

4

4.5 Study 2: Expertise-Driven Alignment for LLM-Powered Psychotherapy Chatbots

Study 1 underscored the value of expert-crafted dialogue scripts for aligning LLMs in psychotherapy. However, as shown in Fig 4.8 and Fig 4.1, the most labor-intensive aspects of expert-scripting involve both creating dialogue content (e.g., questions, reflections, advice) and designing structured dialogue flows. SAG relies heavily on fully expert-authored scripts, which are costly and difficult to scale in expert-driven domains like psychotherapy.

To address this, we introduced Script-Strategy Aligned Generation (SSAG), a more flexible and efficient alignment approach that requires only partial expert input. SSAG drew on evidence-based therapeutic strategies [158, 159] adapted from MI [11, 160] to guide LLMs in generating responses that balance therapeutic adherence with conversational flexibility. SSAG enabled LLMs to co-author therapeutic dialogue content and dynamically manage dialogue flow based on predicted therapeutic strategies, reducing reliance on experts' fully scripted dialogue. In Study 2, we compared SSAG (partial alignment) with LLM-SAG (full alignment) to assess whether SSAG could maintain therapeutic effectiveness while reducing dependence on expert-authored dialogue scripts (RQ2 and RQ3). By streamlining expert input demands and enabling flexible generation, SSAG facilitated more efficient and scalable development of psychotherapy chatbots supported by LLMs.

4.5.1 Approach: Script-Strategy Aligned Generation (SSAG)

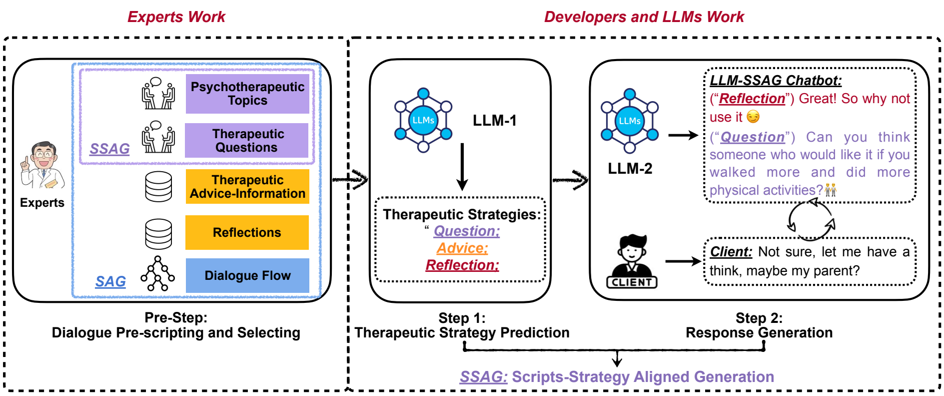


Figure 4.8: Visualization of SSAG and its comparison with SAG. In SSAG, experts provide core therapeutic content, while LLMs first predict therapeutic strategy and then generate responses accordingly. In contrast, SAG relies on fully expert-scripted dialogues.

Script-Strategy Aligned Generation (SSAG) combined partial expert-scripted dialogue content with therapeutic strategies to reduce reliance on fully expert-authored dialogue scripts. As shown in Fig 4.8 and Fig 4.1, expert-crafted dialogue scripts typically include five components: psychotherapeutic topics, therapeutic questions, reflections, and general advice, along with tree-structured dialogue flows. SSAG simplified this by requiring only the three most essential components: psychotherapeutic topics, therapeutic questions, and general advice (optional), as expert pre-authored input. Unlike prior alignment approaches such as unconstrained prompting/fine-tuning that rely heavily on full scripting and limit the therapeutic explainability, SSAG introduced a middle-ground alignment grounded in strategy prediction, offering greater modularity, efficiency, and explainability than direct prompting or fine-tuning alone. Specifically, SSAG operated in the following two steps as illustrated in Fig 4.8:

Step 1: Predicting the therapeutic strategy: SSAG first predicts the next therapeutic strategy based on dialogue context, selecting from core therapeutic behaviors: ”asking questions,” ”reflective listening,” and ”giving advice.”

Step 2: Generating the response: Once the therapeutic strategies are selected, LLM generates a response according to the predicted strategies. For example, if ”asking questions” or ”giving advice” is selected, the LLM retrieves and paraphrases the pre-written expert content (therapeutic questions or advice) based on dialogue context. If ”reflective listening” is selected, the LLM generates reflections freely based on the context, allowing for dynamic, empathetic responses without scripting. By anchoring responses to specific therapeutic strategies, SSAG aligns the conversation with therapeutic principles [158,159], maintaining control over the dialogue flow while enabling flexibility in generation.

4.5.2 Study Methods

Study design and procedure

The procedure of Study 2 is illustrated in Fig 4.2 (b). We conducted a 10-day field study employing a mixed design to compare three chatbot types: (1) a rule-based chatbot (baseline), (2) an LLM-powered chatbot aligned via SAG using prompting (LLM-SAG (Prompt)), and (3) an LLM-powered chatbot aligned via the flexible SSAG approach (LLM-SSAG). Participants were divided into two groups: one interacted with rule-based chatbot and LLM-SAG (Prompt); the other with the rule-based chatbot and LLM-SSAG. Each group followed a counterbalanced design, where participants used one chatbot for five consecutive days, then switched to the other, totaling ten days. After each daily interaction, participants completed evaluation to assess their experience with the chatbot. This design enabled direct comparisons under consistent conditions to evaluate whether SSAG (partial alignment) could deliver psychotherapeutic interventions as effectively as SAG (full alignment).

Types of chatbots

Study 2 examined the following three types of chatbot:

1) **rule-based chatbot.** This chatbot was implemented using the RASA framework [195], the same as the rule-based chatbot used in Study 1 (see Section 4.4.2 (1)).

2) **LLM-SAG (Prompt).** This chatbot was implemented by aligning with the full expert-crafted dialogue scripts using SAG via prompting as in Study 1 (see Section 4.4.2 (4))

3) **LLM-SSAG.** As detailed in Section 4.5.1, SSAG offered a flexible alternative to SAG by reducing the need for fully scripted dialogues. It relied on three core expert-crafted elements: psychotherapeutic topics, a sequence of therapeutic questions, and general advice (optional). LLM-SSAG chatbot operated in two steps (see Fig 4.8). In Step 1, LLM-1 predicted the therapeutic strategies (i.e., asking questions, reflective listening, or giving advice) based on dialogue context. In Step 2, LLM-2 generated the dialogue response, ensuring alignment with the predicted strategies from Step 1.

In this study, SSAG restricted completely "free" generation to reflective responses, while therapeutic questions and advice were retrieved and paraphrased from expert-crafted dialogues. This balanced flexibility and safety: reflections are lower-risk and promote empathy and engagement, while questions and advice, being more directive and sensitive, are anchored in validated expert input to maintain therapeutic integrity and ethical standards. In comparison, SAG places greater emphasis on alignment with expert-crafted dialogues through retrieving and paraphrasing, with either fine-tuning or prompting, to ensure consistency with expert demonstrations across all response types. However, SSAG was designed to be extensible, similar to Retrieval-Augmented Generation (RAG) [216], but with greater controllability over which response types can be freely generated and which must remain grounded in expert data. Although this work restricted question and advice generation anchoring to expert-authored content for safety, future versions could allow more personalized generations with additional validation mechanisms to ensure safety and adherence to ethical standards. This setting also enhanced efficiency and adaptability. Training LLMs to freely generate expert-quality therapeutic questions and advice would require extensive domain-specific data and frequent model retraining. Findings from Study 1 have evidenced that prompting provides a more efficient alignment for expertise integration than fine-tuning. By contrast, SSAG maintained alignment with expert guidance while enabling flexible generation and dynamic dialogue flow management, reducing the need for continuous model retraining and making it more scalable.

The prompt templates used for LLM-powered chatbots are provided in 4.E and

4.F

Measures

To ensure consistency with Study 1, we used the same evaluation metrics, including linguistic quality, engagement, perceived empathy and MI level, therapeutic alliance, and usability. These were assessed daily after each chatbot interaction. Intrinsic motivation to change was measured every two days (pre-, Day 2, Day 4, and post-intervention) using a validated scale adapted from [217].

Participants

Based on a G*Power analysis [207], a minimum of 16 participants was required to detect a medium effect size (d = 0.25) with 80% power and $\alpha = 0.05$. To ensure robustness, 21 participants were recruited (N = 21). All participants were 18 or older, fluent in English, and recruited through institutional channels and social media. They received monetary compensation aligned with the guidelines for completing the participation over 10 days. The study was approved by the institutional ethics committee at the University of Amsterdam. Participant demographics are shown in Table 4.5.

Demographic	**Categories**	**Numbers of Participants (%)**
Gender	Female	13 (61.9%)
	Male	8 (38.1%)
Age	18-24	9 (42.9%)
	25-34	10 (47.6%)
	35-44	1 (4.8%)
	45-54	1 (4.8%)
Education	High school degree or equivalent	2 (9.5%)
	Bachelor’s degree	5 (23.8%)
	Master’s degree	10 (47.6%)
	Doctorate or higher	4 (19.0%)
Professional Domain	Health and Medical Science	3 (14.3%)
	Science, Technology, Engineering, Mathematics (STEM)	5 (23.8%)
	Business, Economics, and Law	3 (14.3%)
	Communication, Arts, Culture and Entertainment	4 (19.0%)
	Education and Social Science	3 (14.3%)
	Government and Public Sector	2 (9.5%)
	Other	1 (4.8%)

Table 4.5: Characteristics of participants in Study 2.

Data analysis

We first checked data suitability using the Shapiro-Wilk test [208] for normality and Bartlett’s test [209] for homogeneity of variance. As data were not fully normally distributed, we applied Generalized Estimating Equations (GEE) [210], a statistical method robust to non-normality, to compare participants’ perception and therapeutic effectiveness across rule-based, LLM-SAG (Prompt), and LLM-SSAG chatbot conditions.

4.5.3 Findings

Pre-validation: automatic evaluation of SSAG

4

Dataset	AnnoMI		BiMISC	
Metric	Accuracy	F1	Accuracy	F1
Flan-T5 [146]	46.2	77.6	19.1	17.4
Vicuna-13B [218]	44.7	76.2	10.5	18.8
GPT-4	**50.0**	**78.9**	**33.6**	**27.9**
GPT-4o (FT)	**63.6**	**81.4**	**47.2**	**36.9**

Table 4.6: The next MI strategy prediction on dataset AnnoMI (single-strategy prediction) and BiMISC (multiple-strategy prediction). ("F1" is the F-score [145])

To support the feasibility of SSAG, we pre-validated it by automatic evaluations. Building on prior work [160, 219], we benchmarked LLMs using two open-sourced MI datasets: AnnoMI [220] and BiMISC [221] for evaluating this prediction task.

First, we evaluated LLMs' ability to predict therapeutic (i.e., MI) strategies (Step 1 of SSAG as defined in Section 4.5.1), which serve as the basis for response generation in Step 2 of SSAG. As shown in Table 4.6, GPT-4 (zero-shot setting) achieved strong performance, while fine-tuning on data further improved accuracy (GPT-4o (FT)). Notably, performance was lower on the BiMISC dataset due to its multi-strategy complexity (i.e., one utterance has multiple MI strategies), compared to the single-strategy in the AnnoMI dataset. Although the overall accuracy was modest, this is acceptable within SSAG, where multiple strategies may be contextually appropriate and no single strategy is universally optimal. These results confirmed that LLMs can support SSAG ' s Step 1 for therapeutic strategy prediction and serve as a basis for guiding dialogue generation in Step 2 of SSAG.

Second, we evaluated LLMs' ability to align with MI strategies for utterance generation (Step 2 of SSAG as defined in Section 4.5.1). We benchmarked several prominent LLMs, focusing on LLMs renowned for their size, performance, and open-source nature. We selected six open-sourced LLMs: **Flan-t5-xxl, Vicuna-13B, Qwen-14B, Qwen2-7B, Llama-2-13B**, and **Llama-3-8B**. All these open-sourced LLMs are recognized for their capability to align closely with human instructions [175], particularly in dialogue interactions. Additionally, we chose **GPT-4** as a commercial benchmark, noted for its superior performance in dialogue generation.

To assess the effect of MI strategy alignment, we compared two prompting types. As shown in Table 4.7, LLMs generated utterances based only on prior dialogue context, without any explicit instruction to follow a MI strategy, denoted as "/wo". Another prompt instructed LLMs for utterance generation with both the dialogue context and the specified MI strategy (e.g., reflection, open-ended question), denoted as "/w". This comparison enabled to evaluate how effectively LLMs can generate responses that adhere to expert-defined therapeutic strategies.

To objectively evaluate the quality of generations, we applied the following automatic evaluation metrics:

- **BLEU & ROUGE** [98,99] assesses the overlap of n-grams between the generation and reference in terms of precision and recall, respectively. We measure $n = 1$.
- **METEOR** [100] evaluates semantic and syntactic accuracy, including synonym and paraphrase use for linguistic precision.
- **BERTScore** [101] assesses semantic similarity by BERT embeddings, measuring contextual relevance of generations.
- **Entropy** [222] quantifies the unpredictability and assesses the effectiveness of strategy in controlling generation. Lower entropy indicates more aligned responses.

Model	Length		BLEU↑		ROUGE↑		METEOR↑		BERTScore↑		Entropy↓	
MI Strategy	/wo	/w	/wo	/w	/wo	/w	/wo	/w	/wo	/w	/wo	w
AnnoMI												
Flan-T5-XXL-11B	13.3	10.9	10.1	10.9	8.2	8.4	10.7	11.6	84.8	**85.6**	**2.9**	**2.6**
Vicuna-13B	40.5	30.1	**14.3**	**14.7**	**12.1**	**12.3**	17.7	17.3	**85.1**	85.5	4.9	4.5
Qwen-14B	38.8	37.7	7.8	12.5	6.5	10.4	10.7	15.1	62.4	84.4	3.8	4.5
Qwen-2-7B	59.7	30.7	11.9	13.8	10.2	10.5	17.3	16.0	84.2	85.2	5.5	4.6
Llama-2-13B	36.2	44.9	7.5	14.5	7.7	11.8	12.0	18.2	79.8	84.4	4.2	5.0
Llama-3-8B	57.6	61.1	8.7	8.7	7.6	8.3	13.8	14.2	81.1	80.7	5.1	5.1
GPT-4	54.3	23.1	13.6	14.3	11.2	11.2	**18.7**	**18.9**	84.1	85.5	5.3	4.3
BiMISC												
Flan-T5-XXL-11B	31.3	26.5	9.5	10.8	9.7	9.7	11.6	12.2	82.7	83.8	**2.7**	**2.9**
Vicuna-13B	51.9	38.0	8.4	12.1	8.0	**10.0**	13.1	16.7	82.7	84.3	4.9	4.7
Qwen-14B	41.2	42.6	7.7	11.1	6.4	9.3	10.7	14.9	61.7	83.9	3.9	4.8
Qwen-2-7B	64.5	40.8	9.1	10.9	7.7	8.3	14.5	14.2	82.1	84.1	5.5	4.8
Llama-2-13B	6.1	20.7	1.6	6.6	4.7	5.2	2.0	7.8	20.3	82.3	4.0	3.9
Llama-3-8B	61.2	61.1	7.2	7.2	7.1	7.4	11.5	11.9	81.2	81.0	5.1	5.0
GPT-4	60.3	36.3	**10.9**	**13.7**	**9.8**	**10.0**	**16.0**	**16.9**	**83.4**	**84.5**	5.4	4.6

Table 4.7: Results from the automatic evaluation on two datasets with seven benchmark LLMs under two different types of prompt: (1) generation without alignment to any specific MI strategy (/wo), and (2) generation explicitly aligned with the given MI strategy (/w). This comparison assesses the impact of strategy-level alignment on the quality of therapeutic dialogue generation.

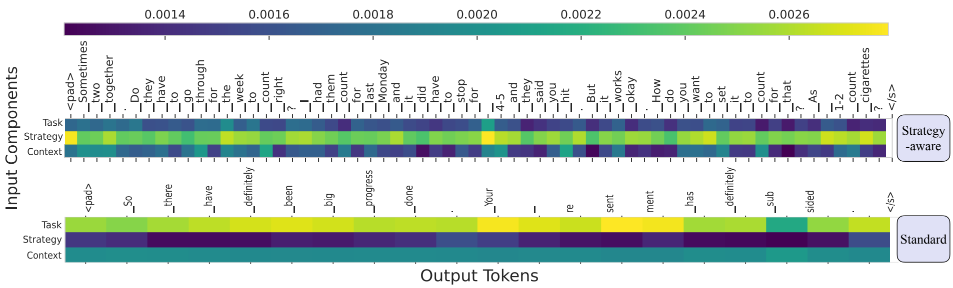


Figure 4.9: Comparison of attention score distributions from LLM (i.e., Flan-T5 [146] with Encoder-Decoder architecture, last layer, and averaged across all heads) for dialogue generation, with ("strategy-aligned") and without ("standard") aligning to the MI strategy. Attention to input tokens was aggregated into three prompting components for better comparison.

Table 4.7 demonstrated that the standard prompt yields the lowest scores in automatic metrics. This outcome showed that strategy-aligned generation with the MI strategy can effectively instruct LLMs to generate dialogue following specific MI principles. From the model perspective, the commercial GPT-4 model consistently achieved the highest scores across metrics. However, notable was the performance of open-sourced LLMs Flan-T5, Vicuna-13B and Qwen2, which closely rivaled that of GPT-4. This highlighted the significant advancements in open-sourced LLMs for MI dialogue generation.

Moreover, to understand how LLMs utilize MI strategies in dialogue generation, Figure 4.9 visualizes the attention distribution [223] of LLM generations with and without MI strategy. The attention distribution for the strategy-aligned generation shows a significantly denser focus on the MI strategy compared to the standard prompting components (i.e., task instruction and dialogue context only). This emphasizes LLMs' consideration of MI strategies in strategy-aligned dialogue generation.

Comparing chatbot types: rule-based vs. LLM-SAG vs. LLM-SSAG

We evaluated three chatbot types: rule-based, LLM-SAG (Prompt), and LLM-SSAG in Study 2 using consistent measures from Study 1. As shown in Table 4.8, LLM-powered chatbots aligned with partial expert-crafted scripts (LLM-SSAG) performed comparably to those aligned with full scripts (LLM-SAG) in both conversational quality and therapeutic effectiveness. Besides, both of LLM-powered chatbots significantly outperformed rule-based chatbots in delivering psychotherapy across various assessing dimensions.

To address RQ2, we compared LLM-SAG (Prompt) and LLM-SSAG to evaluate whether LLM-SSAG chatbot could achieve comparable therapeutic effectiveness and conversational quality to those aligned with full pre-scripted dialogues by SAG. Results from the GEE analysis (Table 4.8) showed no significant differences in performance between LLM-SSAG and LLM-SAG across key metrics, including linguistic quality, dialogue relevance, empathy, engagement, MI adherence, therapeutic alliance, and usability. The findings that no significant performance differences exist between LLM-SSAG and LLM-SAG across most metrics supported the effectiveness of the partial alignment approach in LLM-SSAG, maintaining similar levels of conversational quality and therapeutic adherence while offering a more flexible alternative to the fully aligned LLM-SAG chatbots for psychotherapy. Radar plots in Fig 4.11 explicitly visualized their comparable performances.

As shown in Table 4.8 and Fig 4.10, linguistic quality (LLM-SSAG: M=3.83; LLM-SAG: M=3.84) and dialogue relevance (LLM-SSAG: M=4.09; LLM-SAG: M=4.00) of both LLM-powered chatbots were rated higher than the rule-based chatbot baseline, indicating better conversational quality. Empathy (LLM-SSAG: M=3.70; LLM-SAG: M=3.54), engagement (LLM-SSAG: M=3.23; LLM-SAG: M=3.18), MI adherence (LLM-SSAG: M=3.90; LLM-SAG: M=3.83), and usability (LLM-SSAG: M=4.09; LLM-SAG: M=4.04) were also higher than the rule-based baseline, with LLM-SSAG slightly outperforming LLM-SAG. Both LLM-powered chatbots scored higher in therapeutic alliance (LLM-SSAG: M=3.83; LLM-SAG: M=3.87) than rule-based chatbots, reflecting stronger therapeutic rapport. In terms of the actual motivational change, both LLM-SAG and LLM-SSAG chatbots showed sustained increases in motivation change throughout the intervention period, while the rule-based chatbot showed limited impact on motivation change over time. These results addressed RQ2 by demonstrating that LLM-SSAG chatbot offered comparable performance to the fully aligned LLM-SAG chatbot.

Comparison	Measure	Mean (SD)	Coefficient	Effect ($Std.B$)	p-value
Rule-based vs. LLM-SAG (Prompt)					
	Linguistic Quality	3.26 (.79) vs. 3.84 (.63)	.57	.72 (large)	.01 **
	Dialogue Relevance	3.75 (.70) vs. 4.00 (.42)	.22	.36 (medium)	.18
	Empathy	2.81 (1.13) vs. 3.54 (.84)	.75	.70 (large)	.00 **
	Engagement	2.69 (.89) vs. 3.18 (.98)	.48	.54 (large)	.07
	Perceived MI	3.38 (.74) vs. 3.83 (.52)	.41	.62 (large)	.04 *
	Therapeutic Alliance	3.28 (.65) vs. 3.87 (.52)	.49	.62 (large)	.02 *
	Usability	3.50 (.60) vs. 4.04 (.59)	.51	.71 (large)	.03 *
Rule-based vs. LLM-SSAG					
	Linguistic Quality	3.35 (.83) vs. 3.83 (.73)	.50	.63 (large)	.06 –
	Dialogue Relevance	3.89 (.74) vs. 3.95 (.23)	.23	.37 (medium)	.21
	Empathy	2.67 (.95) vs. 3.70 (1.00)	1.00	.94 (large)	.01 **
	Engagement	2.74 (.69) vs. 3.23 (.93)	.50	.56 (large)	.10
	Perceived MI	3.59 (.69) vs. 3.90 (.56)	.35	.53 (medium)	.09
	Therapeutic Alliance	3.56 (.49) vs. 3.83 (.63)	.37	.53 (large)	.10
	Usability	3.56 (.85) vs. 4.09 (.72)	.56	.78 (large)	.04 *
LLM-SAG (Prompt) vs. LLM-SSAG					
	Linguistic Quality	3.84 (.63) vs. 3.83 (.73)	.01	.01 (small)	.97
	Dialogue Relevance	4.00 (.42) vs. 4.09 (.55)	.09	.15 (small)	.58
	Empathy	3.54 (.84) vs. 3.70 (1.00)	.17	.16 (small)	.63
	Engagement	3.18 (.98) vs. 3.23 (.93)	.05	.06 (small)	.89
	Perceived MI	3.83 (.52) vs. 3.90 (.56)	.07	.10 (small)	.73
	Therapeutic Alliance	3.87 (.52) vs. 3.83 (.63)	.04	.10 (small)	.73
	Usability	4.04 (.59) vs. 4.09 (.72)	.05	.07 (small)	.86
Motivation Change Day-1 vs. Day-5					
	Rule-based	3.74 (.70) vs. 4.07 (.35)	.11	.20 (medium)	.04 *
	LLM-SAG (Prompt)	3.74 (.70) vs. 4.13 (.33)	.39	.67 (large)	.00 **
	LLM-SSAG	3.74 (3.70) vs. 4.14 (.65)	.44	.75 (large)	.00 **

Table 4.8: Results of generalized estimating equations (GEE) [210] comparing rule-based and LLM-powered chatbots via full alignment (SAG) and partial alignment (SSAG). (**p<.01, *p<.05.)

Comparison of reliance on expert-crafted content: full alignment (SAG) vs. partial alignment (SSAG)

To address RQ3, SSAG supported efficient co-development where experts pre-script only essential dialogue elements (i.e., psychotherapeutic topics, key therapeutic questions, health advice (optional)), and LLMs handle response generation or para-

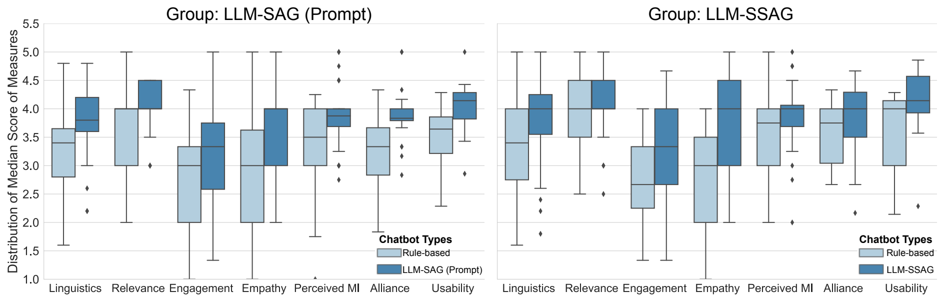


Figure 4.10: Medians of measures in human evaluation across chatbot conditions. Y-axis ranges from 1 to 5 (Likert scale). Each box shows the interquartile range (25th–75th percentile), with the midline indicating the median. Whiskers extend to 1.5 times the interquartile range, and dots outside represent outliers.

phrase as well as dynamic dialogue flow management based on therapeutic strategies and dialogue context, which reduces the need for fully expert-scripted content and structured flows, thereby could lower expert authoring workload.

To quantify this reduction, we compared the number of pre-authored utterances required for SSAG versus SAG. The LLM-SAG (Prompt) chatbot utilized the full expert-crafted dialogue scripts, consisting of 1,876 utterances and complete tree-structured flows (see Fig 4.1). In contrast, LLM-SSAG operated with only 501 therapeutic questions and 203 general advice utterances, covering only 37.5% of the total utterances in the dataset. A detailed breakdown by psychotherapeutic topic is included in 4.A. These results showed that SSAG requires less than two-fifths of the expert-authored dialogue content needed for SAG. While this suggested a potential reduction in expert authoring workload, we did not directly measure time or workload. Thus, these findings should be interpreted as indirect evidence of efficiency, not a conclusive workload evaluation.

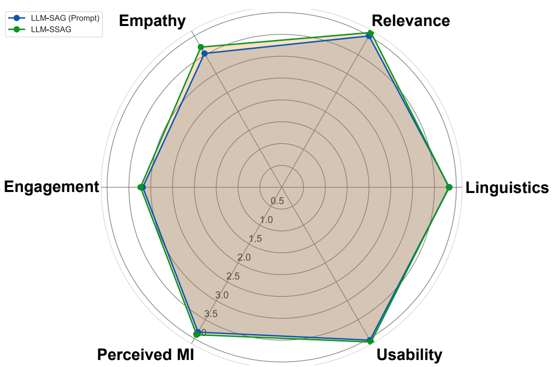


Figure 4.11: Comparison (mean value) between the LLM-powered chatbots employing strict SAG through prompting (LLM-SAG (Prompt)) vs. our proposed flexible alignment (LLM-SSAG). Each axis represents one of six evaluation measures, with mean values plotted for each chatbot type.

4.6 Discussion

4.6.1 Expert-Crafted Dialogues Remain Essential for Chatbot-Delivered Psychotherapy in the Era of LLMs

Chatbots for psychotherapy [4, 7, 147] have traditionally relied on rule-based systems structured around expert-crafted dialogue scripts [13, 14, 165], valued for their safety, controllability, and strict adherence to evidence-based psychotherapeutic techniques like MI and CBT. While reliable, these systems often produce rigid, non-adaptive conversations [12], limiting their ability to tailor responses to nuanced client needs. The emergence of LLMs introduces new possibilities for more personalized, empathetic, and engaging dialogue [155, 168, 224]. However, when deployed without domain alignment, LLMs, such as ChatGPT [65], lack domain knowledge, structure, and safety controls needed for high-stakes contexts like psychotherapy. They struggle to generate sequential, goal-directed therapeutic questions and can deviate from evidence-based psychotherapeutic techniques, making experts in such domains hesitate in relying on them alone.

Study 1 compared four chatbot types: rule-based, pure LLM, and two LLMs aligned with expert-crafted dialogue scripts using fine-tuning or prompting (SAG). Results showed that LLM-aligned chatbots significantly outperformed both rule-based and pure LLMs across assessing metrics. These findings addressed **RQ 1** that *expert-crafted dialogue scripts remain essential in the era of LLMs for effective digital psychotherapy*. Specifically, findings from Study 1 revealed that participants rated LLM-powered chatbots higher than rule-based ones in linguistic quality, empathy, and engagement, highlighting the conversational fluency and personalization that generative models bring. While rule-based chatbots are reliable, their rigidity and reliance on pre-authored scripts limit their ability to adapt to diverse conversational contexts [14]. Participants noted a lack of contextual awareness and personalization, both essential for empathetic and engaging psychotherapy interactions. However, the flexibility of LLMs came with tradeoffs. Pure LLMs were perceived as more natural and engaging but often lacked therapeutic structure and relevance. Participants described them as "human-like" yet vague or unfocused with qualitative insights, resulting in lower scores for dialogue relevance and MI adherence compared to rule-based and aligned LLMs. In contrast, rule-based chatbots maintained therapeutic objectives and structure but were often seen as rigid and less engaging. Aligned LLMs (SAG) effectively bridged this gap by preserving the structure of expert scripts while introducing empathetic and adaptive dialogue generation. Participants appreciated this combination, describing aligned LLM-powered chatbots as "professional yet warm" and capable of delivering structured guidance with emotional resonance, an essential balance in psychotherapy [7, 21]. The qualitative insights reinforced the quantitative findings: aligned LLMs offered a more effective tradeoff between therapeutic rigor and conversational naturalness than either rule-based or unaligned pure LLMs.

Although LLMs are increasingly capable of generating empathetic and engaging dialogue [225, 226], such AI-generated empathy should be interpreted with caution because AI perceives warmth, empathy differently than humans [22]. For instance, LLMs may over-express empathy [158, 227], which can lead to discomfort or unintended emotional effects, and AI often fails to express empathy in positive circumstances while humans do [227]. A study [228] shows conversational agents adapt empathetic responses to certain identities, which can be potentially exploitative. While our findings show that participants perceived LLM-generated responses as more empathetic and engaging than rule-based ones, such empathy and engagement need further control. Prior work suggests that human-AI collaboration can enhance perceived empathy [183], supporting our approach of aligning LLMs with expert guidance to balance conversational attributes (e.g., empathy) with therapeutic structure. Given the relatively short duration of our study, future research should examine how human-perceived chatbot's empathy evolves over longer-term use.

In sum, expert-crafted dialogue scripts remain essential in the era of LLMs, not as rigid blueprints but as scaffolding that enables LLMs to deliver therapeutically-structured, safe, and engaging dialogues. SAG-aligned LLMs guided by the expert-crafted scripts outperformed both rule-based and pure LLMs by effectively managing the tradeoff, offering a practical path for engaging and therapeutically effective chatbot-delivered psychotherapeutic interventions.

4.6.2 Efficient Alignment of LLMs with Domain Expertise for Psychotherapy

In psychotherapy, aligning LLMs with expert knowledge is essential to ensure conversations remain therapeutically relevant and effective. Prior work on instructing and aligning LLMs [21, 155, 160] has focused primarily on two alignment approaches: fine-tuning and prompting, each with distinct trade-offs. Fine-tuning [186] embeds domain knowledge directly into model parameters but is resource-intensive and requires large, curated datasets, often scarce in sensitive fields like psychotherapy. In contrast, prompting [187] offers a cost-effective and flexible alternative by guiding LLMs at inference time using structured inputs, avoiding retraining while encoding therapeutic goals through prompt design. Prompting has been shown to use less than one-tenth the computational cost of fine-tuning for achieving similar alignment [197], making it a more scalable solution for psychotherapy where domain-specific data is limited but adherence to therapeutic principles is critical. It also allows quick adaptation to new therapeutic techniques, supporting flexible deployment in evolving psychotherapeutic settings.

To explore RQ1.2, Study 1 compared these two alignment approaches for their effectiveness in delivering health interventions using the same expert-crafted dialogue scripts. The findings revealed that the prompted chatbot consistently outperformed the one by fine-tuning across assessing measures. Prompting preserved the structure of expert-authored dialogue flows while offering greater adaptability in response generation, better supporting context-aware, engaging, and goal-directed therapeutic conversations. This difference in our work partly arises from how each approach handles dialogue structure: fine-tuning embeds the tree structure into the model, reducing adaptability, while prompting uses tree-structured scripts [191] during inference, enabling more flexible, diverse conversational paths while preserving therapeutic coherence and staying on the scripted therapeutic path, as noted by our participants, "It (LLM-SAG (Prompt)) can understand my specific questions and give correct answers, then it can go back to previous talk after my questions, interesting." This tradeoff is particularly important in psychotherapy, as prior work [21, 184] indicates that adherence to therapeutic principles should not come at the expense of empathy and conversational engagement.

However, both alignment approaches (i.e., SAG through either prompting and fine-tuning) rely heavily on extensive expert input, making scalability and efficiency a challenge. We thereby proposed Script-Strategy Aligned Generation (SSAG), a more flexible alignment that reduces scripting overhead by leveraging only core components (i.e., therapeutic topics, key questions, and optional advice) while allowing LLMs to support the dialogue generation and manage dialogue flow dynamically through therapeutic strategy prediction, inspired by prior CSCW research that uses therapeutic strategies to enhance human counseling effectiveness [158, 159] as well as NLP studies [160, 184] that instruct LLMs to follow evidence-based therapeutic principles. Results from Study 2 showed that LLM-SSAG performed comparably to LLM-SAG (Prompt) across most evaluation metrics, addressing our RQ2. This demonstrated that SSAG can maintain therapeutic structure by anchoring each response to a predicted MI strategy (e.g., asking a question, offering advice, or providing a reflection), ensuring alignment with evidence-based therapeutic behaviors while allowing flexible generation. By guiding LLMs through strategy prediction, especially reflective listening, SSAG tried to mitigate the over-reliance on therapeutic "method" (e.g., CBT) seen in prior work [21] and supports more engaging, empathetic dialogues. This structured-yet-adaptive design mirrors techniques like Retrieval-Augmented

4

Generation (RAG) [216], but with greater controllability for expert oversight.

The choice of MI and CBT balanced LLMs by combining MI's emphasis on empathy and engagement with CBT's structured, goal-oriented interventions. However, SSAG is not limited to these two focused frameworks. Its stepwise, strategy-level design allows integration of other therapeutic techniques by substituting proper expert-defined strategies. Because SSAG operates at the level of therapeutic strategies rather than fixed language patterns, it is also adaptable to multilingual and culturally diverse contexts by accommodating different language norms and counseling styles.

4.6.3 Implication: LLM-Supported Co-Development of Psychotherapy Chatbots

Psychotherapy chatbots, as socially impactful tools, hold promise for expanding access to mental health support, particularly in underserved communities. However, their development has traditionally relied on labor-intensive expert scripting, which limits scalability. Our work addressed this challenge by first evaluating alignment approaches and then introducing Script-Strategy Aligned Generation (SSAG), a method designed to reduce dependence on fully expert-scripted dialogues and enable more efficient co-development of scalable and explainable LLM-powered psychotherapy chatbots between AI (i.e., LLM), developers, and domain experts.

From a technical perspective, Study 1 confirmed that prompting is a more scalable and resource-efficient alignment method than fine-tuning. Prompting preserved therapeutic adherence while allowing for flexible, human-like interaction, making it suitable for conversational therapies such as MI and CBT, where structured intervention must be delivered through adaptable dialogue. Building on this, SSAG introduced a two-stage strategy-level alignment method that relies on partial expert input (i.e., therapeutic topics, key questions, and general advice), while delegating dialogue generation and flow management to the LLM via predicted MI strategies. Results from Study 2 showed that SSAG matched the performance of fully aligned SAG chatbots across key therapeutic metrics while requiring less than 40% of expert-authored dialogues, offering strong evidence of reduced reliance on expert-crafted dialogue content, answering our RQ3. While we did not directly measure reductions in expert workload by SSAG, its design eliminates the need for domain experts to manually construct full dialogue trees or write structured prompts. This suggests potential for easing expert involvement during the chatbot development process. For instance, experts can focus on defining high-level strategies and therapeutical questions, while the LLM handles context-specific generation, reducing the need for technical prompt engineering. We acknowledge that this only partially demonstrates collaborative development. While SSAG facilitates division of labor between domain experts and AI, the degree of collaboration was not empirically studied.

Future CSCW research should more directly evaluate expert-AI workflows using participatory design methods, task-based workload assessments, or longitudinal co-development trials to rigorously assess reductions in human workload of both expert and developer.

From a CSCW and human-AI collaboration perspective, SSAG advances the vision of cooperative development in expert-driven and sensitive domains. SSAG positions LLMs as flexible yet controllable collaborators that support experts in constructing adaptive, therapeutically grounded dialogues, operating within the bounds of expert-guided therapeutic strategies. This draws on prior CSCW work on human-AI co-creation frameworks [229] and domain-specific chatbot co-development [152], demonstrating how structured human-AI collaboration can scale innovation in mental health technologies. Although we evaluated it in a semi-controlled setting, deployment in unsupervised settings still introduces ethical and safety concerns. Real-world deployment requires additional safeguards, including human-in-the-loop oversight [230, 231], real-time validation mechanisms, and adherence to ethical standards. These measures are especially critical in unsupervised or high-risk settings like psychotherapy.

Taken together, SSAG aligns with CSCW's mission to foster computer-supported cooperative work in socially impactful domains. By reducing development barriers and enabling scalable, explainable, and expert-aligned AI, SSAG opens pathways for delivering cost-effective and accessible mental health support at scale, particularly in underserved or resource-limited fields. Its design facilitates a scalable and efficient development between experts and developers, illustrating how AI can be harnessed in ways that are explainable, domain-aligned, and co-developed. Through this lens, this work serves as both a technical advancement and a CSCW-relevant contribution, representing a promising step toward building cooperative, expert-informed and socially impactful AI technologies that advance mental health support systems and the broader goals of CSCW.

4.7 Limitations and Future Work

Despite the contributions of this work, we acknowledge several limitations that remain for future exploration.

First, while the expert-crafted dialogue scripts used in this study are valuable, the reliance on such data introduces potential biases, which could impact the quality of LLM alignment. Its limited size and sole focus on MI and CBT, may not fully capture the diversity of real-world therapeutic interactions or generalize well to other more intensive forms of psychotherapy that require different conversational structures and techniques. Future work should expand datasets to cover a wider range of therapeutic scenarios enhancing SSAG's generalizability and robustness. Moreover, although SSAG demonstrated reduced reliance on full scripting, requiring fewer than 40% of expert-authored utterances compared to SAG, we did not directly measure expert workload or time savings. Thus, the efficiency claim should be interpreted as indirect. Future work should include task-based workload assessments or participatory design studies to evaluate the reduction of human workload more rigorously.

Second, although Study 2 included a 10-day field evaluation, the sample size was modest. Given the personalized nature of psychotherapy, user evaluations may vary widely and the reliance on self-reported feedback limits the generalizability of our findings. Future work should involve more diverse samples and incorporate expert assessments for more robust evaluations. While the study duration was relatively short, prior HCI and digital mental health research has employed comparable intervention periods (e.g., 2 weeks) to assess the effectiveness of such tools or applications [232–234]. Nonetheless, human-perceived AI empathy and warmth may evolve over time. Longitudinal studies are needed to examine how these perceptions and therapeutic outcomes develop with continued use in real-world deployments.

Third, despite efforts to ensure control and explainability, deploying LLM-powered chatbots in psychotherapy raises ongoing ethical and safety concerns. Even when aligned with expert-crafted dialogues, LLMs may produce unpredictable or inappropriate responses that pose risks in sensitive mental health contexts. Addressing these concerns requires safeguards such as real-time monitoring, fallback mechanisms for high-risk situations, and stronger data privacy and explainability protocols.

Fourth, participants' prior exposure to chatbots or similar AI-driven services may have influenced their perceptions and interactions. Those with more experience might have had different expectations or levels of comfort, potentially affecting their evaluations of the chatbots. While we did not explicitly control the prior exposure for participant recruitment, the randomized and counterbalanced study design helps mitigate systematic bias. Future work could explore this variable more directly with stratified analyses or larger samples to better understand how prior exposure shapes user perceptions in chatbots in psychotherapeutic contexts.

While this work advances the alignment of LLMs with expert-crafted dialogues, addressing these limitations is essential to ensure the scalability, generalizability and ethical deployment of LLM-powered psychotherapy chatbots.

4.8 Conclusion

Our work provided initial evidence that aligning LLMs with expert-crafted dialogues may significantly enhance their performance and user perceptions in chatbot-delivered psychotherapy for behavioral interventions. By creating a dataset of expert-crafted dialogue scripts for Motivational Interviewing (MI) and Cognitive Behavioral Therapy (CBT), we conducted empirical studies comparing rule-based, pure LLM, and LLM-aligned chatbots to evaluate their conversational quality and therapeutic effectiveness. Our proposed approach, Script-Strategy Aligned Generation (SSAG), integrated human expertise with LLMs, resulting in controllable and engaging chatbots. This approach mitigated data scarcity, reduced reliance on expert input, and enhanced the controllability of LLMs in delivering psychotherapy for behavioral interventions. These findings underscored the importance of human-guided LLMs for controllable and explainable digital mental health tools, and contributed to the field of CSCW by advancing scalable, cost-effective and LLM-supported development of psychotherapy chatbots, a promising path for future research in digital mental healthcare and health behavioral intervention.

Chapter appendix

4.A Overview of the Expert-Crafted Dialogue Scripts

We provide a detailed overview of the dataset we created in this work and introduced in Section 4.3.

Technique	Psychotherapeutic Topic	Description of Each Topic	Num of Question / Advice / Total Utterance per Topic
Overall			501 / 203 / 1876
MI	Greeting	Establishes rapport and initiates a dialogue.	2 / 0 / 9
	Transition topic	Facilitates smooth shifts between therapeutic topics.	8 / 0 / 18
	General introduction	Introduces the purpose of the conversation.	14 / 11 / 54
	Status of physical activity	Explores current levels of physical activity to understand baseline behavior.	15 / 13 / 81
	Handling Bad Weather	Discusses strategies to maintain motivation during poor weather.	14 / 14 / 65
	Sleep Quality	Gathers information about sleep habits and explores links to energy/motivation.	3 / 12 / 32
	Setting a Goal	Helps the user identify a meaningful, self-directed behavioral goal.	8 / 4 / 36
	Setting a Goal follow up	Revisits previously set goals to reflect on progress, challenges, or adjustments.	27 / 2 / 118
	Motivational onboarding for energy	Educates and motivates the user to consider movement as a way to boost energy.	11 / 9 / 50
	Rating confidence of physical activity	Uses confidence scales to explore and enhance self-efficacy for being active.	10 / 11 / 68
	Evoking self-efficacy	Encourages the user to recognize their own strengths and capacity for change.	10 / 16 / 103
	Improving motivation	Identifies personal reasons for change to boost intrinsic motivation.	39 / 10 / 117

Table 4.A.1: Overview of expert-crafted dialogue scripts. "Num of Question / Advice / Total Utterance per Topic" indicates the number of therapeutic questions, general advice, and total utterances included in the dialogue script for each psychotherapeutic topic.

Technique	Psychotherapeutic Topic	Description of Each Topic	Num of Question / Advice / Total Utterance per Topic
Overall			501 / 203 / 1876
CBT	Supportive social environment	Encourages reflection on positive social support for behavior change.	15 / 7 / 101
	Supportive social environment follow up	Assesses ongoing support and identifies ways to strengthen helpful connections.	22 / 5 / 79
	Mindfulness	Introduces present-moment awareness as a way to manage thoughts and emotions.	17 / 5 / 52
	Mindfulness follow up	Reviews practice experience and reinforces benefits of mindfulness.	7 / 8 / 27
	All or nothing thinking	Identifies black-and-white thinking patterns that can block progress.	47 / 1 / 122
	All or nothing thinking follow up	Applies alternative thinking strategies to reduce rigid thought patterns.	22 / 2 / 63
	Should-statements	Identifies unhelpful internal rules (e.g., “I should exercise every day”).	31 / 5 / 73
	Should-statements follow up	Challenges and reframes unrealistic expectations or self-criticism.	15 / 1 / 49
	Implement Intentions	Helps translate intentions into action using planning and cue-based strategies.	45 / 11 / 130
	Stress management	Teaches techniques (e.g., reframing, breathing) to cope with stress effectively.	42 / 9 / 107
	Habit building	Supports creation of consistent, repeatable behaviors that lead to long-term motivation and change.	22 / 17 / 103
	Time management	Offers tools to prioritize and schedule healthy activities realistically.	18 / 12 / 78
	Unsupportive social environment	Identifies social barriers and develops strategies to handle them constructively.	25 / 13 / 91
	Small wins	Encourages small, achievable actions to build momentum and confidence.	12 / 5 / 50

Table 4.A.2: Overview of expert-crafted dialogue scripts. ”Num of Question / Advice / Total Utterance per Topic” indicates the number of therapeutic questions, general advice, and total utterances included in the dialogue script for each psychotherapeutic topic.

4.B Example of Expert-Crafted Dialogue in MI

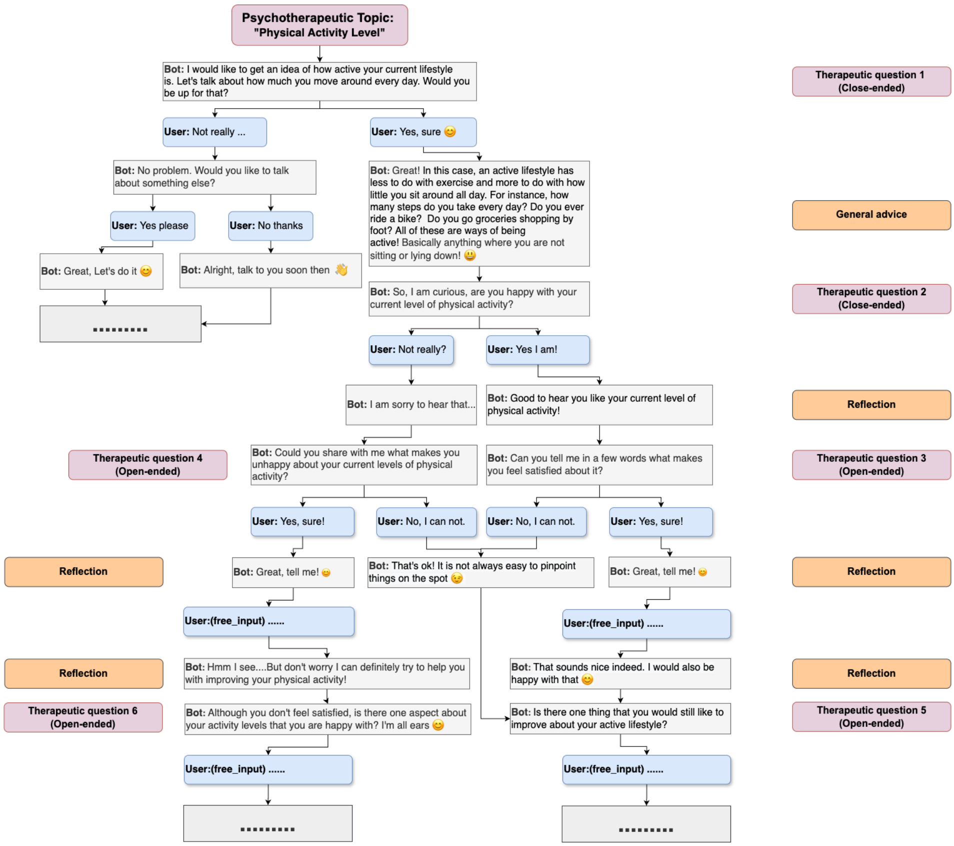


Figure 4.B.1: Examples of expert-crafted dialogues in motivational interviewing (MI) with psychotherapeutic topic "physical activity level". The expert-crafted dialogues contain 1) psychotherapeutic topics, 2) therapeutic questions (both open-/close-ended questions), 3) reflections, 4) general advice, 5) dialogue flows pre-designed by experts.

4.C Example of Expert-Crafted Dialogue in CBT

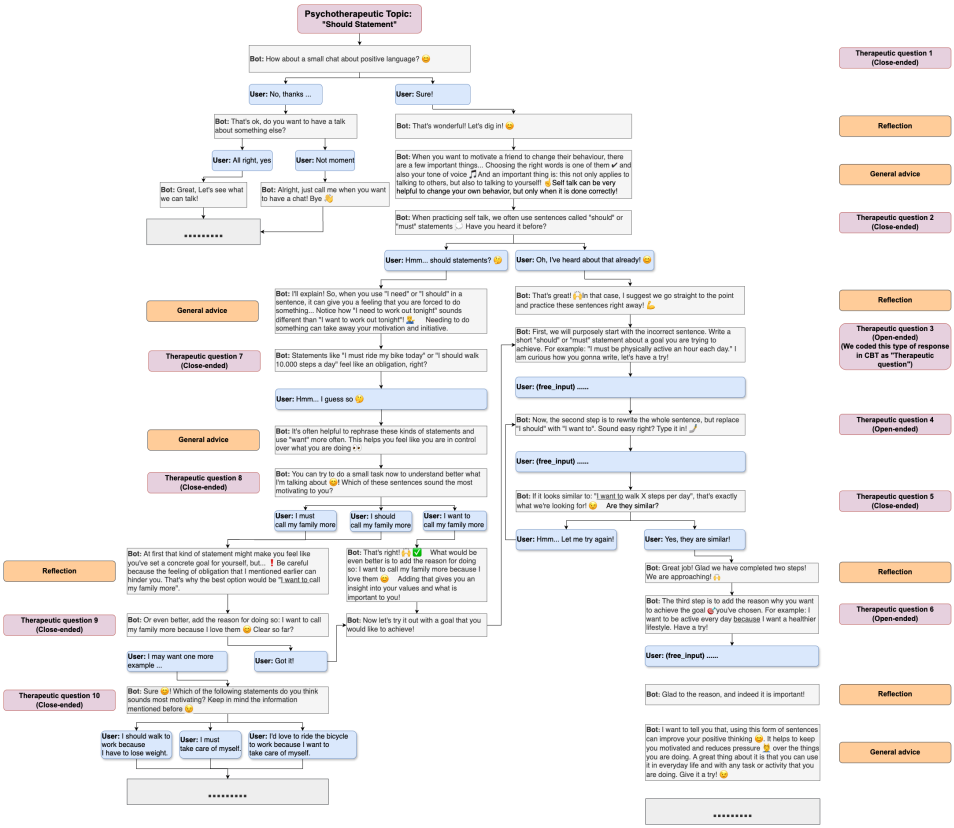


Figure 4.C.1: Examples of expert-crafted dialogues in cognitive behavioral therapy (CBT) with psychotherapeutic topic ”should statement”. The expert-crafted dialogues contain 1) psychotherapeutic topics, 2) therapeutic questions (both open-/close-ended questions), 3) reflections, 4) general advice, 5) dialogue flows pre-designed by experts.

4.D Prompt Template Used in Study: Pure LLM & LLM-SAG (FT)

Prompt Components	Content
Conversation context	**[The continuously accumulated context of the conversation]** Conversation context: Therapist: [...] Client: [...] Therapist: [...] Client: [...] [......]
Task instruction	**[The base instructions to explain the generation task]** Task: You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy). The current therapeutic topic is [given topic]. The definition of this topic is [description of the given topic].

Table 4.D.1: The prompt template used in the Study 1 for pure LLM chatbots and LLM-SAG (FT) chatbots.

4.E Prompt Template Used in Study: LLM-SAG (Prompt)

Prompt Components	Content
Conversation context	**[The continuously accumulated context of the conversation]** Conversation context: Therapist: [...] Client: [...] Therapist: [...] Client: [...] [......]
Expert-crafted dialogue script	**[The expert-crafted dialogue script (including both dialogue content and flow) for the specific psychotherapeutic topic of the current session.]** All dialogue flows in expert-crafted dialogue scripts of [given topic]: Flow 1: [Pairs of Therapist-Client Dialogues] Flow 2: [Pairs of Therapist-Client Dialogues] Flow 3: [Pairs of Therapist-Client Dialogues] Flow 4: [Pairs of Therapist-Client Dialogues] [......]
Task instruction	**[The base instructions to explain the generation task]** Task: You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy) following tree-structured dialogue flows. The current therapeutic topic is [given topic]. The definition of this topic is [description of the given topic]. Rules to follow: 1. Identify the relevant dialogue tree based on the current conversation context and user input. 2. At each conversational turn, please use the Tree-of-Thought technique about the decision point and use Breadth-First Search to go level-by-level in the dialogue tree. Do not explore deeper branches unless the current path is chosen. 3. If no exact match is found, please intelligently adapt while staying close to the current dialogue flow/tree. 4. You can give tailored responses to user input and provide dialogue diversity, but please keep staying strictly to the given dialogue flow/tree. 5. Follow the dialogue tree structure and generate possible next responses based on relevance and coherence and with a clear format. 6. If the ongoing conversational topic is finished, please say goodbye to the user and ask the user to go click the button on the left to evaluate this chatbot.

Table 4.E.1: The prompt template used in Study 1 and 2 for the LLM-SAG (Prompt) chatbots.

4.F Prompt Template Used in Study: LLM-SSAG

Step	Prompt Components	Content
Step 1 of SSAG		
	Conversation context	**[The continuously accumulated context of the conversation]** Conversation context: Therapist: [...] Client: [...] Therapist: [...] [......]
	Therapeutic strategy	**[The definition of the therapeutic strategy]** The definition of the therapeutic strategy: "reflection": [...] "question": [...] "advice": [...]
	Task instruction	**[The base instructions to explain the generation task]** Task: As a psychotherapist of Motivational Interviewing, please predict the next appropriate therapeutic strategy only from the set ["reflection", "question", "advice"] based on the conversation context. The next therapeutic strategy is (are):
Step 2 of SSAG		
	Conversation context	**[The continuously accumulated context of the conversation]** Conversation context: Therapist: [...] Client: [...] Therapist: [...] [......]
	Next therapeutic strategy	**[The next therapeutic strategy of the chatbot response]** The next therapeutic strategy is (or strategies are): "Reflection" or "Question" or "Advice" or their combination based on the prediction from Step 1. The definition of the therapeutic strategy: [definition of the predicted therapeutic strategy].
	Task instruction	**[The base instructions to explain the generation task]** Task: You are a psychotherapist conducting a session to promote healthier behavior using Motivational Interviewing (or Cognitive Behavioral Therapy). The current therapeutic topic is [given topic]. The definition of this topic is [description of the given topic]. You should generate the response strictly aligned with the next therapeutic strategy above.

Table 4.F.1: The prompt template used in Study 2 for the LLM-SSAG chatbots.

4.G Implementation Details of Generations with LLMs

We utilized the August 2024 edition of GPT-4o, coded as `gpt-4o-2024-08-06` [1]. We used `openai` Python library to generate with GPT-4. We opted for default hyperparameters, including the temperature as default to control the randomness of generation. The models were used in compliance with their respective licenses and terms at the time of the study. OpenAI provides a Terms of Use[2].

[1] `https://platform.openai.com/docs/models/gpt-4o`

[2] `https://openai.com/policies/terms-of-use`

Part II

“Interface”: Understanding Human Trust Perception in LLM-Powered Health Information Seeking

5

From Agents to Interfaces: Understanding Trust in Health Information from LLM-Powered Conversational Search

This chapter is based on the following publications:

Authors: Xin Sun, Yunjie Liu, Jos A. Bosch, and Zhuying Li.

Original title: Interface Matters: Exploring Trust Perception in Health Information from Large Language Models via Text, Speech, and Embodiment

Published in: (Journal) Proceedings of the ACM on Human-Computer Interaction (Track Computer-Supported Cooperative Work And Social Computing CSCW)

Authors: Xin Sun, Rongjun Ma, Xiaochang Zhao, Janne Lindqvist, Jan de Wit, Zhuying Li, Abdallah El Ali, Jos A. Bosch.
Original title: From Agents to Interfaces: Understanding Trust in Health Information from Conversational Search
Submitted to: Journal of Behaviour & Information Technology (Under review)

Abstract

Large Language Models (LLMs) deployed through Conversational User Interfaces (CUIs) are transforming health information-seeking by offering immediate, interactive experiences compared to traditional search engines like Google. However, how trust may be influenced by both the types of search agents and the interface used to disseminate the information remains underexplored. This research integrated two mixed-methods studies (lab sessions and interviews) to comprehensively explore trust perceptions in health information across different search agents and dissemination interfaces. In Study 1 (N=21), we conducted a within-subjects lab experiment to compare trust in health information retrieved from ChatGPT and Google across three types of health-related search tasks. The findings indicated that participants generally perceived ChatGPT's responses as more trustworthy than those from Google. Study 2 (N=20) extended the investigation to explore how the dissemination interface influenced trust in LLM-sourced health information by comparing three interfaces: text-based, speech-based, and embodied, all sourcing from the same LLM. Participants' trust ratings varied significantly depending on the interface. Interviews from both studies revealed key factors influencing trust in LLM-powered conversational search, including source credibility, participants' search autonomy, and prior knowledge as well as the interaction style and interface modality. These findings showed that information source and interface play a key role in trust in health information and suggested a preference for LLM-powered conversational search to transform health information-seeking. The results underscored the interplay between the credible search agents and the thoughtfully designed dissemination interfaces in shaping trust. These insights are crucial for developing effective, trustworthy LLM-powered health tools that enhance the health information-seeking experience.

5.1 Introduction

THE RELIANCE on online sources for health-related information has grown significantly due to their convenience and rapid access. Historically, search engines like Google have been the main tools for individuals seeking health information on diseases, treatments, and general wellness advice such as diet and exercise [23, 24, 235, 236]. However, the emergence of generative AI has introduced Large Language Model (LLM)-powered conversational agents like ChatGPT as increasingly popular alternatives [237]. These LLM-powered agents engage users in interactive dialogues, transforming information-seeking into a personalized and conversational experience [238]. This evolution is particularly significant in health information contexts, where the ease of access, accuracy, presentation and trustworthiness of information directly affect health decisions and outcomes [239–242]. These advancements in search technologies highlight the need for renewed perspectives on how these search agents reshape people's trust, specifically trust in health information presented. Mayer et al.'s model [41] defines trust as a willingness to be vulnerable based on expectations of competence, benevolence, and integrity. In human-AI interaction, Lee and See [43] extend this to viewing trust as an attitude that an agent will help achieve one' s goals under uncertainty. In health contexts, trust also entails a sense of safety in using the information for decisions that may impact well-being. Work from Sillence et al. [24] reveals that trust is essential in adopting online health advice, influencing how users evaluate information and decide whether to act on it. While extensive research has investigated trust in traditional search engines like Google, the factors influencing trust in LLM-powered conversational search remain underexplored. With the rise of LLMs, people are no longer just searching for web-listed information; they are engaging in dynamic conversations, and factors shaping trust are now more dynamic, complex, and overall multifaceted [243–245].

LLM can be accessed through various user interfaces that likewise shape user trust [36, 246]. Conversational user interfaces (CUIs), such as text-based chatbots and voice assistants act as essential links between users and health resources. These interfaces enable direct communication with health information and facilitate dialogue between healthcare professionals and patients [247]. Trust in the information can vary depending on the interaction style and modality, such as text, speech, or embodied interactions [30, 248]. Key factors influencing trust in AI-driven health interactions include the source of information, presentation style, and the design of user interfaces that deliver the content. For instance, a trustworthy source presenting health advice in a clear style within an intuitive interface may foster greater confidence in

5

the information provided. Embodied interfaces, which simulate advanced human-like traits, can enhance engagement but may also raise concerns about authenticity, privacy, and the "uncanny valley" effect [249–251]. Thus, it remains unclear how LLM-powered dissemination interfaces affect public trust in health information searches, highlighting a gap in understanding their role in health-related information searches.

To better understand the complexity of trust in health-related information, this work focused on two perspectives: search agents and dissemination interfaces. It investigated how users perceive information sourced from LLM-powered agents compared to traditional search engines. A better understanding of these perceptions can inform the design of more effective, user-centered digital health tools, where both the search process and interface design reinforce trust in online health information. Conversely, this work also aligned with the need to recognize how interface or content features, some of which can subconsciously be interpreted as trust cues, may (un)intentionally be misused, leading to misplaced trust. Designing with such awareness can help prevent unintended harms and promote more responsible health communication.

Given the foregoing, this work seeks to address the following three research questions:

- **(RQ1)** Does people's perceived trust in health-related information differ (a) between traditional search agents (e.g., Google) and LLM-powered conversational agents (e.g., ChatGPT)? (b) when the information is delivered through different dissemination interfaces (i.e., text-based, speech-based, or embodied)?
- **(RQ2)** Is trust in such retrieved online health information correlated with the inherent trust in specific (a) search agents and (b) dissemination interfaces?
- **(RQ3)** Which factors contribute to perceived trust in personal health information across different: (a) search agents and (b) dissemination interfaces?

To address these research questions, we adopted a mixed-method approach, conducting two studies aimed at understanding trust in health information from the perspectives of information search agents and dissemination interfaces. The procedure of these two studies is demonstrated in Figure 5.1. In Study 1, we conducted a within-subject lab study (N=21) in which participants completed three health-related information search tasks using two different search agents: Google and ChatGPT. Trust was measured at three points—before, between tasks, and after the lab session. Follow-up interviews were conducted to explore the factors influencing participants' trust perceptions. In Study 2, we examined how different dissemination interfaces, specifically text-based, speech-based, and embodied CUIs, influence trust in the information presented. We conducted a within-subject lab study (N=20) where participants interacted with each of the three interfaces to complete three types of health-related search tasks. By observing and analyzing participants' interactions, we aimed to understand how the presentation of information via different CUIs impacts trust. Follow-up interviews were also conducted to further explore participants' motivations behind their trust perceptions.

Several interesting findings emerged from our studies. First, the results revealed significant differences in trust influenced by search agents (Google and ChatGPT) and dissemination interfaces (text-based, speech-based, and embodied CUIs). The conversational and personalized features in LLM-powered conversational search significantly enhanced trust compared to Google's more conventional list-based presentation. Besides, dissemination interfaces played a crucial role, with text-based CUI generally perceived as the most trustworthy due to its ease of use and participants' familiarity level. Second, we identified common factors contributing to trust variations across both studies, including prior experience, information presentation style, and the interactive mechanism with LLM-powered tools. However, some factors differed by context. For instance, in Study 1, human-like features were highly valued and positively influenced trust. In Study 2, the introduction of more advanced human-like features, such as verbal interactions and embodied interfaces, enhanced engagement but raised concerns about privacy and the authenticity of interactions, ultimately affecting trust. Third, we proposed implications based on both studies, suggesting that future research should balance anthropomorphic and personalized features with usability while considering the specific context and modality of health information dissemination. This balance helps foster trust and enhances user experience in digital health tools.

Our work provides empirical evidence on how people perceive trust in AI-generated health information during online searches. As reliance on online health information grows, ensuring the trustworthiness of these eHealth tools is considered crucial. By examining how participants perceive trust in LLM-powered conversational agents compared to traditional search engines within dissemination interfaces, we highlight the critical role of design, specifically usability, familiarity, and interaction modality, in shaping trust perceptions in retrieved and delivered health information. Our findings suggest practical guidelines for developing trustworthy and user-centered eHealth tools, emphasizing the enhancements of reliability, transparency, and engagement of digital health information search. These insights may inform future research on designing and evaluating trust-building strategies in LLM-powered digital health tools, ensuring that such tools are reliable, user-centered, and capable of effectively addressing diverse user needs in the evolving landscape of generative AI.

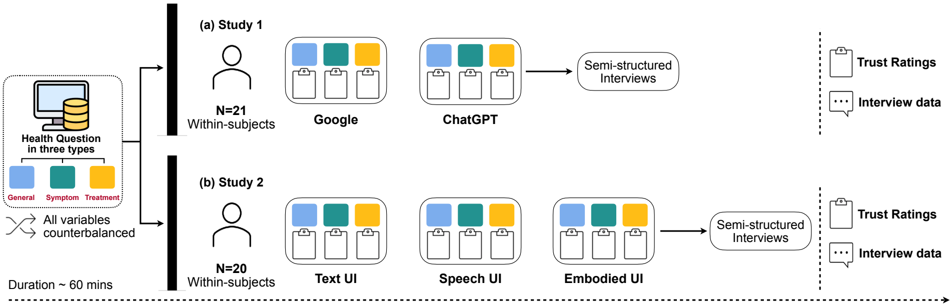


Figure 5.1: Visual summary of the studies in this paper. (a) Study 1: Lab study to explore the trust perception in health information using two search agents: Google vs. ChatGPT; (b) Study 2: Lab study to investigate perceived trust in health information disseminated by three different conversational user interfaces: text-based, speech-based, and embodied interface.

5.2 Related work

5.2.1 Theoretical background of trust in health context

Trust is a multifaceted psychological construct central to both interpersonal and human-technology interactions [43]. Mayer, Davis, and Schoorman's integrative model [41] defines trust as a willingness to be vulnerable to another party based on the expectation that the party is competent, has benevolent intentions, and acts with integrity. In the context of human-AI interaction, Lee and See [43] extend this to automation, defining trust as "an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Building on this, McKnight et al. [252,253] highlight that trust in technology is shaped by several factors, including perceived usefulness, ease of use, reliability, credibility, and potential risk involved.

In healthcare contexts, trust becomes even more critical due to the sensitive nature of health information and its potential impact on health decision-making [37,38]. Here, trust is not simply about believing the information provided but also about feeling safe to rely on it for meaningful decisions that can affect well-being. The continuous evolution of healthcare, driven by technological innovation, demands that trust-building mechanisms also adapt to changing user perceptions and technological capabilities. Understanding how trust forms in digital health environments requires both theoretical grounding and empirical validation. Work from Johnson et al. [254] identifies three core dimensions that shape trust in online health information: (1) source credibility, (2) content relevance and reliability, and (3) system design and usability. These dimensions directly informed the design of the work by guiding our focus on (1) comparing trust across different source agents (i.e., Google vs. ChatGPT) to assess source credibility, (2) examining perceived trust in the content provided across tasks to reflect content relevance, and (3) evaluating how different user interfaces (text-based, speech-based, and embodied) impact usability and overall trust perceptions in LLM-powered health information.

Research by others further confirm that these three dimensions of trust identified by Johnson et al. [254] collectively affect users' overall trust perceptions in health information. Several studies emphasize the importance of source credibility as a cornerstone of trust [24, 46, 48, 244, 255–258]. Additionally, the design and usability of health tools strongly influence trust perceptions, with user-friendly and well-structured systems being more trusted than those with complex or unclear designs [119, 121, 125, 259, 260]. User-specific factors, such as prior experience and familiarity level with the system [261], as well as pre-existing expectations [262], also affect trust. In addition, the quality of health information itself [263] is essential for fostering trust. Research has highlighted factors such as the credibility [121,125,140,141,143,264], clarity [254], and relevance [257,261,262,265] of information as key contributors to user trust. Furthermore, transparency in data handling, privacy protection, and adherence to ethical standards have emerged as critical determinants of trust in health contexts [170, 266–270].

5.2.2 Trust in health information through web-based search

Online health information searching has become a main way for people to get health advice due to its convenience [23, 271, 272]. Over the past decades, web-based health information search usually relied on search engines (like Google) and some medical websites [235, 236, 273]. Search engines, with their powerful retrieval capabilities, can offer users a wide range of health information. People can quickly obtain relevant health information by simply entering related keywords, which is why search engines like Google have become the primary tool for most people to retrieve information [23, 274], including health-related information. Additionally, some professional online medical websites, like NIH [273] and Mayo Clinic [236], aim to provide authoritative and reliable health information and advice, serving as sources of health information online. The growing reliance on online sources for health advice has prompted extensive research into how people perceive and trust online health information and its sources [124, 125, 140, 143, 243–245, 259].

Past research [24, 47, 256, 275] has investigated how people foster trust in online health information and websites, providing valuable insights into the process of trust formation. This body of work has been crucial in helping us understand more systematically and scientifically. Not surprisingly, many existing studies emphasize that the reliability of the information source is very important in this context [46, 48, 243, 244, 255–258, 276]. Information from well-known medical institutions, peer-reviewed scientific journals, or recognized healthcare professionals is more likely to be trusted. On the other hand, websites with user-generated content or commercial motives are often questioned [140, 259, 271, 277, 278]. Alongside source credibility, the intrinsic quality of the information itself also plays a crucial role [121,140,141,259,277,279]. Moreover, how the information is presented also affects its trustworthiness [119,121,125,259]. For instance, a well-designed, easy-to-navigate website or app is usually seen as more trustworthy than a platform with a confusing or complicated user interface. Research also shows that users tend to trust search engines' ranking algorithms, often perceiving higher-ranked results as more credible [124, 280]. However, concerns about the potential for search engine manipulation and dissemination of misinformation have arisen [269, 270, 281]. Emerging concerns about data privacy and ethical considerations are also increasingly influencing trust levels [266, 267, 269]. Transparency regarding how personal health data is stored, used, and protected can have a significant impact on human trust [170, 267, 268].

5.2.3 Trust in health information through LLM-powered conversational search

Conversational search [29, 237] steps into the field of health, with ChatGPT (and similar commercially available LLM-based chatbot services) demonstrate capability in handling health-related queries, although research in this area is still in its early stages. A recent systematic review [239] highlighted as advantages of using ChatGPT in healthcare settings by offering automated patient consultations, preliminary diagnoses, and general health recommendations [240, 241]. However, there are also concerns about the application of ChatGPT in healthcare. Ethical issues are raised regarding the potential biases of the information and the protection of sensitive health data [239,282]. Hristidis [283] compared ChatGPT and Google for queries related to dementia and other forms of cognitive decline. Results showed that ChatGPT was rated as more objective and relevant, while Google offered more up-to-dateness and reliability. It was not determined how these outcomes determined overall trust. Another study [284] found that medical experts considered responses from ChatGPT as more trustworthy, with 40 percent rating ChatGPT's responses as more valuable than Google's. However, all responses from the above studies were assessed by medical service providers and researchers. It is not yet clear whether the general public, with more limited domain knowledge, would perceive health-related information differently. Furthermore, most previous studies focused on the quality of the information obtained from ChatGPT and Google but not on how this impacted trust perceptions.

Unlike the extensive literature on search engines like Google, the factors influencing people's trust in LLM itself and its responses remain largely unexplored. Studies on other conversational agents such as chatbots, have identified multiple potential determinats of trust, such as the communication style, the depth of medical knowledge displayed, and chatbot "personality" [285]. Studies also suggest that the trustworthiness of an LLM may be influenced by its consistency in offering health information [286] and its transparency [287]. Overall, although conversational search by LLM shows great potential in the health sector, further research is necessary to understand the trust people have in its information. It is also relevant to compare trust levels between conversational search agents and traditional search engines like Google for health information searches.

5.2.4 Trust in conversational user interfaces for health information seeking

As conversational user interfaces (CUIs) become more prevalent in healthcare, research has focused on factors influencing perceived trust in web-based interfaces, such as source reliability [46, 48, 243, 244, 255–258, 274], information credibility [121, 141, 143, 259, 277], quality [121, 140, 245, 279], and presentation styles [119, 121, 124, 125, 259].

Building on the foundations of trust in web-based interfaces, Mayer, Davis, and Schoorman's trust model [41] highlights the importance of credibility (competence), user-centric intentions (benevolence), and ethical standards (integrity) in CUIs. Trust in CUIs for health information seeking varies and influenced by multiple factors such as individual user characteristics (e.g., prior experience) and interface features (e.g., usability, modality, and design). As might be expected, the credibility and relevance of information are key to perceived trust [254, 274]. In addition, interface design, including information presentation and accessibility, and usability [288], also play critical roles in fostering trust. Persuasive design principles such as social presence and perceived expertise can also enhance trust [289]. Luger et al. [290] additionally emphasize the importance of CUIs exhibiting intelligent behavior and contextual understanding to build user trust. Personalizing interactions to individual preferences can increase trust by enhancing perceived system intelligence. Consistency with users' expectations [262, 290] and prior experiences [290] also reinforces trust, as does the credibility of information sources [48, 243, 244, 255]. Recently, multi-modal CUIs that utilize text, voice, and visual cues to deliver information present unique challenges for trust formation. While these multi-modal interactions can enhance user engagement and understanding [291], maintaining consistent reliability across modalities is crucial. Moreover, characteristics of voice and visual representation can also influence trust [292–295]. Therefore, providing consistent and reliable information, and interaction across modalities is essential to sustain trust.

With the advancement of CUIs with multi-modal capabilities, significant ethical and privacy challenges emerge, impacting user trust. Transparency about data storage, usage and protection is vital for user security [266, 296, 297]. The collection and use of sensitive health data require stringent data protection. Additionally, the potential misuse of CUIs, such as spreading harmful or misleading information [16], raises critical ethical concerns, particularly in health context. Thus, robust ethical guidelines and data protection policies are necessary to ensure CUIs are trustworthy and ethical tools for health information seeking. As these interfaces evolve, ongoing research is essential to address the challenges of trust and harness the full potential of LLM-powered CUIs in the health contexts.

5.3 Study 1: Comparison of search agents for health information seeking

We conducted a mixed-methods study consisting of a controlled experimental lab study followed by semi-structured interviews. In the lab study, we employed a within-subjects design to examine how people interact with two different search agents: the search engine Google and the LLM-powered agent ChatGPT.

We conducted a mixed-methods study consisting of a controlled experimental lab study followed by semi-structured interviews. In the lab study, we employed a within-subjects design to examine how people interact with two different search agents: the traditional search engine Google and the LLM-powered conversational agent ChatGPT.

5.3.1 Study Methods

Participants

A power analysis using G*Power [207] indicated that a minimum of 20 participants were needed to detect a medium effect size with an alpha level of .05 and 80% power. Accordingly, we recruited 21 participants (N=21) through the institute's recruitment system at the University of Amsterdam. Participation was voluntary and informed consent was obtained before the lab sessions. Each participant received a monetary compensation according to the institute's guidelines. Participants were required to be fluent in English and experienced in online information searching. The study received approval from the ethics and data protection committee of the University of Amsterdam (ID: FMG-2336-2023)

The characteristics of participants are summarized in Table 5.1. Most participants (71.4%) were between 18 and 24 years old. Regarding education, 47% held undergraduate degrees, 47.6% had postgraduate qualifications, and 4.7% held doctorate degrees. Participants represented various academic fields: 47.6% from social sciences, 19% from business and commerce, 14.3% from computer and information technology, and 19.1% from other sectors. In terms of online health information seeking, 42.8% frequently used online sources, 52.3% occasionally, and 4.7% rarely or never relied on online resources.

Demographic	**Categories**	**Numbers of Participants (%)**
Gender		(N=21)
	Female	16 (76.1%)
	Male	5 (23.9%)
Age		
	18-24	15 (71.4%)
	25-34	5 (23.8%)
	65+	1 (4.8%)
Education		
	High school	1 (4.7%)
	Bachelor	9 (42.9%)
	Master	10 (47.6%)
	Doctor	1 (4.7%)
Professional Domain		
	Social Science	10 (47.6%)
	Business and Commerce	4 (19.0%)
	Health and Medical Science	1 (4.7%)
	Computer Science & Information Technology	3 (14.3%)
	Other	3 (14.3%)
Frequency of online health information seeking		
	Often	9 (42.8%)
	Sometimes	11 (52.3%)
	Rarely	1 (4.7%)
Frequently used search agent		
	Search engine	21 (100%)
	Conversational agents	15 (71.4%)
	Social media platforms	8 (38.0%)

Table 5.1: Characteristics of participants in Study 1.

Search agents

We used two search agents for health information in this study, as shown in Figure 5.2: Google and ChatGPT. The study was conducted in the institute's experimental behavioral research laboratory facility in a quiet room with desktop computers running the Windows operating system. Participants accessed both search agents via Chrome browser (version 117.0.5938.92) by visiting their official websites. For interactions with ChatGPT, participants used the GPT-4 model (without web browsing capability). This setup ensured all participants had a consistent experience while using up-to-date technology for each search agent.

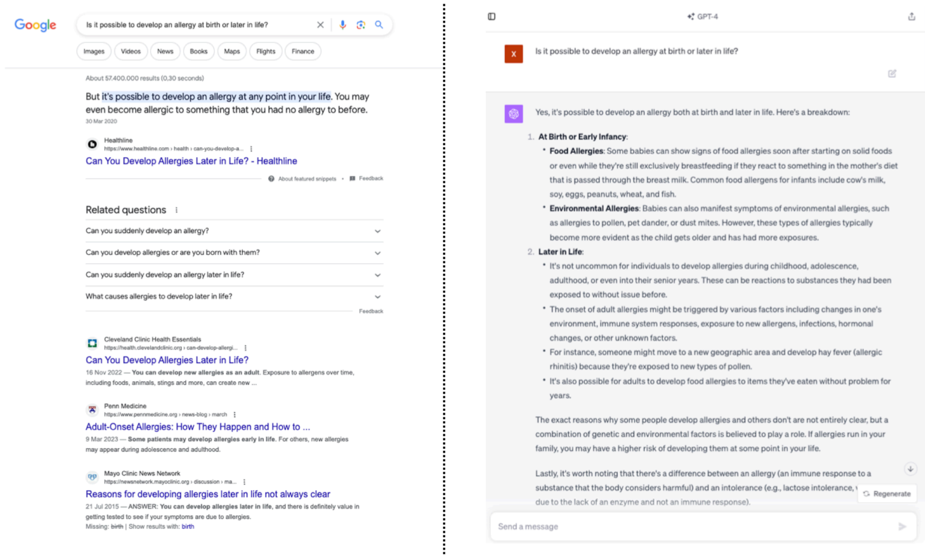


Figure 5.2: Interface of search agents with a search task example using the first type of search task (informational health questions): 'Is it possible to develop an allergy at birth or later in life?'

Search tasks

In this study, search tasks were health-related questions that we asked participants to find answers to during the lab session. Each participant completed six search tasks, selected from an open-sourced dataset [298] with individual health questions from Yahoo [299], which contains personal health-related questions labeled by types. We pre-selected 25 questions from each of the three types to ensure diversity and comprehensiveness. The complete list of questions used in the study is included in Appendix 5.A.1. Three types of personal health-related questions are described as follows:

General health questions. This group of questions aims to gather general knowledge or facts about a specific health topic. Examples include: *"Do you have information about Weight Control?"*; *"Do you have information about Vitamin D?"*

Symptom and cause-related health questions. This group of questions revolves around understanding symptoms or causes associated with particular health conditions. The example questions are: *"What are the symptoms of the eating disorder?"*; *"What causes Memory loss?"*

Treatment-related health questions. This group of questions seeks information about potential treatments for specific conditions. Examples include: *"What are the treatments for dry eye syndrome?"*; *"How can I lower my heart rate?"*

Measures

Before the lab study, we inquired about demographic information and their usage experience with different search agents. During the study, we measured participants' trust in health-related information obtained from the two search agents. Below, we detail the specific measurements used.

Propensity of trust in technology. To gauge participants' inherent trust in technology, we administered a pre-study questionnaire [300] that assessed their general tendency to trust technology. This scale consists of six items rated on a 5-point Likert scale (from 1=Strongly Disagree to 5=Strongly Agree). An example item was: *"I think it ' s a good idea to rely on technology for help."*

Human perceived trust in health-related information. After each search task, participants rated their trust in the retrieved health information. We used the validated 'Trust of Online Health Information' questionnaire [254, 301] to assess the credibility, reliability, and believability of information. This survey included eleven items rated on a 5-point Likert scale (from 1 = Strongly disagree to 5 = Strongly agree) with a Cronbach's alpha of .95 in our study. An example item was: *"The information appears to be objective."*

Human perceived trust in the search agent. Participants rated their trust in search agents' ability to deliver health information after completing all three search tasks with each agent. The 'Model of Online Trust of Health Information Websites' questionnaire [274], consisting of 15 items on a 5-point Likert scale (from 1 = Strongly disagree to 5 = Strongly agree) (Cronbach's alpha=.62) was used to assess trust in agents. An example item was: *"The search agent (e.g., Google) provides truthful information."*

Intention to use search agent for health-related information searching. We measured participants' intention to use the specific search agent for health information with a single item on a 5-point Likert scale (from 1 = Strongly disagree to 5 = Strongly agree). The item was phrased as follows: *"I would intend to use Google (or ChatGPT) for health-related information seeking."*

Study procedure

The overview of Study 1 procedure is outlined in Figure 5.1. Before the lab session, all participants provided informed consent according to the institute's guidelines. We employed a within-subjects design with two conditions: searching with Google and searching with ChatGPT. Each participant completed three distinct search tasks using one search agent and then another three different tasks using the other agent, with task orders and agent usage counterbalanced to mitigate order bias. Participants were allowed to interact freely with the assigned search agent until they found a satisfying answer. After each search task, we asked participants to rate their trust in information via a brief survey. Upon completing all tasks with each search agent, participants rated their overall trust in the agent itself.

Following the lab, we conducted semi-structured interviews lasting about 25 minutes to identify participants' interactions with the agents and their reasoning behind their trust ratings. The interviews began with general questions about participants' daily use of different tools for online health information search, followed by specific questions about their experiences with each search agent, such as *"How do you validate the information you receive from Google?"* and *"How do you feel about the answers retrieved from ChatGPT?"* The interviews concluded with a discussion comparing their experiences with both agents and considerations for potential improvements. The same question phrasing was used for Google and about ChatGPT to enable comparison of the answers. The interview protocol used in Study 1 is included in Appendix 5.B.1.

Data analysis

We collected both quantitative and qualitative data to understand how people interact with two different search agents for health information and assess trust in the health information obtained from these agents.

For the quantitative analysis, we first confirmed the robustness of our data by conducting the Shapiro-Wilk test [208] to check for normality and Bartlett's test [209] to assess variance homogeneity, ensuring our data met the assumptions needed for further analysis. We then used two-way repeated-measures ANOVA [302] and paired samples t-test [303] to identify differences in trust perceptions between the agents, alongside performing correlation analyses to explore variable relationships.

We conducted a thematic analysis [304] of the qualitative interview data. Initially, the first three authors transcribed the interviews to familiarize themselves with the data. Subsequently, three coders open-coded the same transcript independently to develop an initial codebook that guided further coding. With the agreed-upon codebook, three coders divided and independently coded the remaining interviews. Codes were iteratively refined and expanded as new insights emerged. We did not seek interrater reliability (IRR) because we viewed coding as a method to explore open-ended research questions about trust factors, rather than as the primary output of our research (see recommendation by [305]). The coders met regularly to review their work and ensure consistency in their coding. Together, they discussed themes relevant to our research questions, identifying patterns that emerged across the interviews.

	Google Mean (SD)	ChatGPT Mean (SD)
Propensity of trust in technology (PPT)	3.87 (.33)	
Trust (search task 1)	3.69 (.90)	4.01 (.74)
Trust (search task 2)	3.81 (.89)	4.10 (.74)
Trust (search task 3)	3.82 (.88)	4.03 (.81)
Trust (Avg of three tasks)	3.77 (.64)	4.05 (.47)
Trust in search agent	3.31 (.99)	3.58 (.98)
Intention to use	3.57 (.49)	3.52 (.85)

Table 5.2: Descriptive statistics of variables in Study 1, while the statistical results are shown in Figure 5.3. Search task 1 is General health questions, search task 2 is Treatment-related questions, and search task 3 is Symptoms and Diagnosis-related questions.

5.3.2 Quantitative Findings: Lab Sessions

Descriptive statistics

We measured participants' general trust in technology, their trust in health information provided by two search agents, and their trust in the agents themselves. The descriptive results are presented in Table 5.2. The average propensity of trust in technology among participants was moderately high, indicating a general positive trust towards technology. Trust in health information varied between search agents; participants generally trusted information from ChatGPT more than from Google. Trust in the search agents also reflected a preference for ChatGPT over Google.

Human trust in health-related information differs by search agents

We conducted a two-way repeated-measures ANOVA to compare differences in trust between two search agents. The analysis was performed to explore if there are differences in trust, including trust in the retrieved information and search agents.

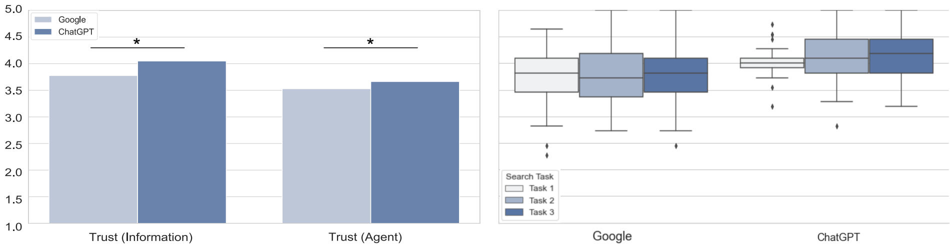


Figure 5.3: Left: Mean of trust scores (Y-axis shows 1–5 Likert scale) in health information and in search agent. Horizontal lines with asterisks denote significant differences between agents (*p<.05). Right: Mean of trust scores in health information for each search tasks separately. Each box represents interquartile range (IQR), the horizontal line inside indicates the median, and whiskers extend to 1.5×IQR. Individual dots represent outliers.

As illustrated in Figure 5.3, the analysis revealed a significant main effect of the search agent on trust in health-related information, $F(1, 20) = 6.73, p = .017$, and $\eta^2 = .057$. This suggests a notable difference in trust, with ChatGPT generally being trusted more than Google for health-related advice. Trust levels did not show significant changes across different types of search tasks, $F(2, 40) = 0.63, p = .480$, with $\eta^2 = .006$. This indicates that the type of health question did not significantly influence trust in the retrieved information. The interaction between search agents and task types was not significant, $F(2, 40) = 0.20, p = .777$, and $\eta^2 = .002$. This indicates that trust differences between the agents were consistent across various search tasks, confirming the stability of trust across different types of search tasks.

Additionally, we did a paired sample t-test to explore the human perceived trust in the search agent itself. The trust scores in Google and ChatGPT as search agents were statistically significant with the difference in means as 0.27, t(21)=-2.53, p=.02, Cohen's d=0.55 (see in Figure 5.3).

Correlation of trust across search agents

We conducted a Pearson correlation analysis to explore the relationships between measured variables, detailed in Figure 5.4.

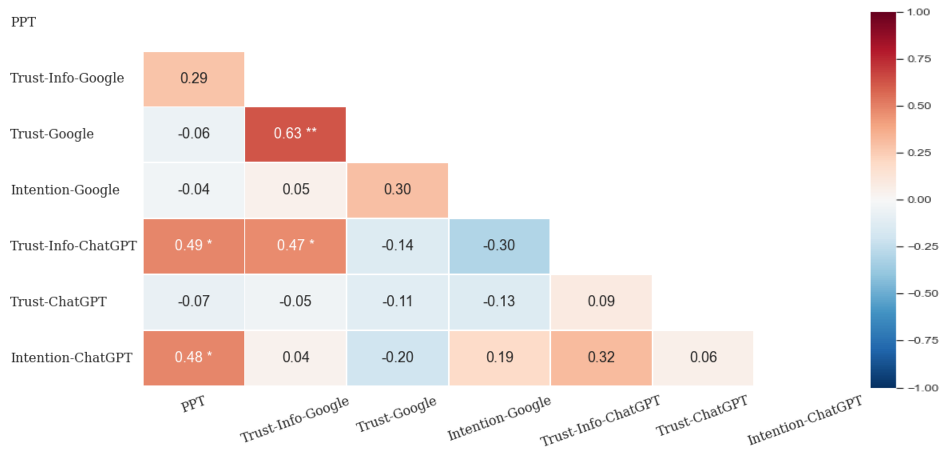


Figure 5.4: Pearson correlation with measured variables, including propensity to trust in technology (PPT), trust in health information (Trust-Info-), and agents (Trust-), and intention to use the agent for further health information searches (Intention-). Significant correlations are marked (**p<.01, *p<.05).

Significant findings included a positive correlation between trust in health information from Google and trust in Google as a search agent, indicating that higher trust in information correlated with higher trust in the agent itself. For ChatGPT, while trust in its health information also positively correlated with trust in ChatGPT, it was not statistically significant. Additionally, there was a notable positive correlation between general trust in technology and trust in health information from ChatGPT, but this correlation did not extend to trust in ChatGPT as a search agent. No significant correlations were found between general trust in technology and trust in Google, either as a source of information or as a search agent. Moreover, intention to use ChatGPT for health information correlated positively with the propensity of trust in technology, a correlation absent with Google.

5.3.3 Qualitative Findings: Semi-Structured Interviews

In this section, we report on themes generated from interview data to understand the formation of trust. The findings are organized into two sections: *How do people search for health-related information* provides a contextual overview of how individuals utilize online sources to find health-related information and offers insights into the scenarios in which trust is shaped. *What factors influence trust in information* illustrates the factors that influence trust in information retrieved from two different agents.

How do people search for health-related information?

Searching online is a pre-step to deal with health-related problems. Although our participants regularly search online for various health-related topics, none face significant health concerns. The major topics for online searching include *diet and exercise* (P2, P14-15, P17, P19, P21), accident treatment such as *tore a muscle* (P2), personal health conditions such as *allergies* (P6), and *mental health issues* (P4). Since most participants considered health-related information serious, they approached information searching cautiously, withholding a degree of trust because *"Everything I see on the Internet isn't necessarily true"* (P4). Participants typically didn't expect to find a perfect solution when searching online. Instead, they aimed to refresh their existing knowledge, learn general information about the topic, or gain direction for the next step. After searching, participants selectively adopted online advice depending on the importance and urgency of health matters, as well as the practicality and risks of the advice.

> *"So if it says like it take a pill or something like that, I won't take it. But if says like, drink water, stay in bed, of course, I'll do that." - P10*

People use their own combinations of sources for health-related information. People have their own preferences for health information sources, including forums, social media, and certain websites. Participants recognized the credibility and practicality of specific resources based on their experience. For instance, many participants trusted official web domains like the website of national health organizations and Google Scholar (P1-2, P4-12, P14-21). Some participants naturally trusted big brands like Google (P7). Others considered social media reliable because influencers' demonstrations help them match symptoms more practically (P11, P17, P19). Additionally, people do not rely on a single resource but a combination of multiple sources. *"Double-check if another source says the same, make a comparison"* (P9) helps participants validate the information. To incorporate ChatGPT into the landscape of information sources, participants suggested a scenario where ChatGPT is used as a preliminary step before conducting searches on Google:

> *"But also if I didn't know anything about a health issue, I could look it up on the ChatGPT first to also like get a sense of direction. What this could be related to and then continue my search on Google." - P1*

What factors influence trust in information?

Previous experience of using an agent influences people's trust. Judgment on the trustworthiness of retrieved information largely depends on participants' prior knowledge of the topic, which helps them assess the information. However, several factors differentiating ChatGPT from Google also influence trust. Previous positive and negative experiences influence people's current trust in search agents. Positive experiences of recovering after following online advice increase trust in that source (P11), whereas negative experiences with misinformation create uncertainty:

> *"ChatGPT can give me some results really quickly, but if I copy the DOI number and double-check it, you'll find they're totally wrong. [...] So that's one of the important reasons I don't fully trust ChatGPT." - P8*

Compared to Google, ChatGPT is a new tool with which most participants have limited experience. While people are still figuring out how to search for health-related information with ChatGPT, their trust may be influenced by *"news and rumors"* (P21). On the other hand, many participants use Google by default, but trust is weakened by experiences like excessive commercial information (P1-2, P4, P10-11, P13, P16, P18-19, P21) and *"filter bubbles"* (P7) that isolate information to personalize searches [306].

Information presentation influences people's trust. Our participants believed that presenting information in a professional yet understandable form is crucial for trust. This includes language expression, information structure, and visual cues. First, is the language expression. Our participants believed the medical field is highly specialized and health information often contains specific terminology. They think that professional use of language and terminology enhances the credibility of the information but also makes it harder for them to understand the topics they're searching for (P1, P2, P17). Besides terminology, the tone of language expression also affects trust, such as the use of passive voice and confidence in expression. The passive voice helps convey objectivity, leading to greater trust (P17). Furthermore, the confidence with which

ChatGPT delivers information significantly impacts participants' trust. An uncertain answer that includes terms like "*maybe*" or "*I'm not sure*" (P10) lowers trust. However, an overly certain answer can also appear untrustworthy:

> *"if it's too, too confident in stating this and this leads to this [...], (I'll) trust less, Because the answer it's never so straightforward." - P16*

Other factors, such as information structure and visual cues, also play a role in shaping trust. A well-structured response guides participants to understand the result, and a logical flow enhances the reliability of the information (P11). In the context of Google, visual cues, such as being "very colorful" (P14) or having "unprofessional logos on the websites" (P4), can lead to lower trust regardless of the content of the information.

Autonomy influences trust In most cases, control of the search process increases people's certainty and trust in the answer. Google provides more control, giving participants autonomy in each step. Access to rich information enhances trust because "*with Google you have more the feeling that you have your own judge*"(P13). However, this method is not always preferred, especially when lacking knowledge for evaluating the information.

> *"The more information you see, the more overwhelming it can also get, so you may find out you don't know everything from googling too much. And then ChatGPT typically can make it easier and give you (information) more trustworthy." - P16*

On the other hand, our participants shared that if they have enough domain knowledge to evaluate answers from ChatGPT, a straightforward and highly relevant answer from ChatGPT is preferred (P1, P5, P7, P9-11).

Human-like interaction influences trust The major difference between searching with Google and ChatGPT lies in the interaction. The way participants interacted with agents influences their trust, and preferences for these interactions vary depending on the situation. Many participants compared interacting with ChatGPT to talking with a human. The human-like dialogue with ChatGPT allows participants to follow up on specific details and receive personalized responses. This context-aware process enhances trust, particularly in personal situations, such as when participants were trying to understand potential causes of pain in specific parts of the body (P11).

> *"You can just ask it (ChatGPT) like you would ask your friend and it understands what you're talking about." - P2*

Personalization influences trust When discussing improvements to AI tools that could increase trust, participants frequently mentioned personalization. Health information varies across regions due to differences in medical systems between countries (P2, P20) and at an individual level based on personal symptoms or allergies (P14). To increase trust, our participants want ChatGPT to tailor its answers by considering these regional and individual differences. For example, act as a mediator to facilitate appointments with doctors, and proactively inquire about personal situations like a human doctor (P2, P14).

> *"(ChatGPT) will ask you some relevant questions about your symptoms and maybe then, it can also find more specific information for you. Maybe it's a different story if you're a man and you're 80 years old, than if you're a woman and you're 18 years old like that, maybe it would give you different information." - P14*

5.4 Study 2: Comparison of dissemination interfaces for health information

We employed a mixed-methods approach including lab sessions and semi-structured interviews. A within-subjects design was used to examine how participants interact with different LLM-powered CUIs for health information seeking.

5.4.1 Study Methods

Participants

A power analysis using GPower [207] determined that at least 15 participants were needed to detect a medium effect size with an alpha level of .05 and 80% power. We recruited 20 participants (N=20) through the institute's recruitment system. Participants had to be fluent in English, and participation was voluntary. Informed consent was obtained prior to the lab sessions. Participants received compensation in accordance with institute guidelines. The study received approval from ethics and data protection committee of the University of Amsterdam (ID: FMG-8695-2024).

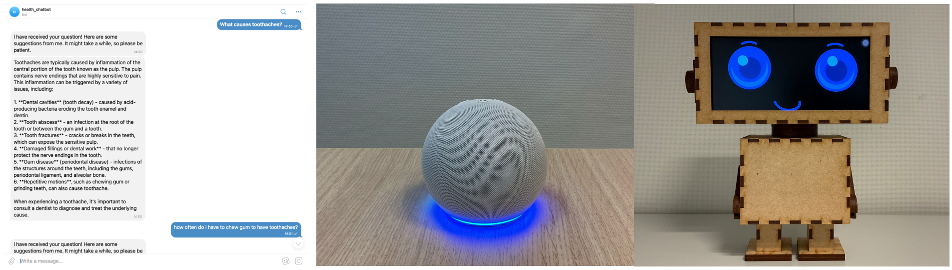
The demographic information of the participants is listed in Table 5.3. Participants were aged between 18 and 34 years, with 80% falling in the 18-24 age bracket. Participants came from a variety of professional fields: 40% from Business, Economics, and Law, 25% from Arts, Culture, and Entertainment, 10% from Health and Medical Sciences, 10% from STEM, and 15% from other fields. Regarding educational backgrounds, 55% had undergraduate degrees, and 45% held postgraduate qualifications. As for online health information-seeking experience, 10% frequently used online sources, 75% occasionally searched online, and 15% rarely or never used online resources.

Demographic	Categories	Numbers of Participants (%)
Gender		(N=20)
	Female	16 (80%)
	Male	4 (20%)
Age		
	18-24	16 (80%)
	25-34	4 (20%)
Education		
	Bachelor	11 (55%)
	Master	9 (45%)
Professional Domain		
	Health and Medical Science	2 (10%)
	Business, Economics, and Law	8 (40%)
	Arts, Culture and Entertainment	5 (25%)
	Science, Technology, Engineering, and Mathematics	2 (10%)
	Other	3 (15%)
Frequency of online health information seeking		
	Often	2 (10%)
	Sometimes	15 (75%)
	Rarely	3 (15%)

Table 5.3: Characteristic of participants in Study 2.

Dissemination interfaces

As shown in Figure 5.5, Study 2 utilized three distinct LLM-powered conversational user interfaces: a text-based, a speech-based, and an embodied interface.



(a) text-based interface *(b) speech-based interface* *(c) embodied interface*

Figure 5.5: Three dissemination interfaces were used in the lab study for participants to complete the health-related search tasks.

Text-Based Interface: We developed a web-based chat interface utilizing the GPT-4 model [83] to realize straightforward, text-only interactions. This setup aimed to assess user trust in information provided solely through text.

Speech-Based Interface: Verbal interactions were enabled through an Amazon Echo Dot, also powered by the GPT-4 model, allowing users to engage in natural, spoken conversations and ask follow-up questions. The wake word like 'Hi Alexa' was eliminated. The device displayed a 'hello' indicator only during the initial activation, with a light at the bottom serving as the sole interaction indicator.

Embodied Interface: This setup featured a self-built body, incorporating a mobile phone to display simple virtual facial expressions as the "face". This embodied interface featured a physical, body-shape presence, enhancing the interaction with a tangible form. A small speaker placed inside the body served as the output for vocal responses, connected to the GPT-4 model to ensure consistency across interfaces. This interface was designed to investigate the influence of a physical presence on user trust. The embodiment interface employed basic expressions to focus solely on the effect of physical presence on trust, avoiding biases from advanced features like human likeness or varied expressions.

Search tasks

The search tasks in this study were designed to align with those used in Study 1 to ensure consistency which is introduced in Section 5.3.1. The tasks included 75 questions, categorized into three distinct types of health questions: general, symptom, and treatment-related questions.

Measures

To ensure consistency, the measures used in this study were aligned with those in Study 1. Additionally, we assessed participants' background using validated instruments before the lab study. AI literacy was measured using an 11-item scale [307] on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree), assessing participants' familiarity and comfort with AI (e.g., "I can distinguish if I interact with an AI or a real human"). eHealth literacy was measured using the 8-item eHEALS scale [308], also on a 5-point Likert scale, measuring participants' ability to find, assess, and use online health information (e.g., "I know how to find helpful health resources on the Internet"). During the lab study, perceived usability was evaluated using two items from the UMUX-Lite scale [309], again on a 5-point Likert scale (e.g., "The interface meets my requirements for health information seeking"). Intention to use each interface was measured with a custom single item adapted from TAM3 [310] on a 5-point Likert scale: "Assuming I have access to a [text-based/speech-based/embodied] interface, I would intend to use it for health information seeking."

Study procedure

The overview procedure of Study 2 is illustrated in Figure 5.1. Prior to the lab session, all participants provided consent in accordance with the institute's guidelines. The study involved evaluating interactions with three different interfaces: text-based, speech-based, and embodied. To prevent bias, these interfaces were presented in a counterbalanced order. Each participant engaged in three search tasks per interface, resulting in a total of nine tasks across the three interfaces. Participants were allowed to interact freely with each interface, including asking follow-up questions until they were satisfied with the answers. After completing each search task, participants rated how much they trusted the information they received. Once all three tasks for a specific interface were finished, participants provided an overall assessment of that interface, focusing on trust, usability, and their intention to use it in the future.

Each lab session ended with a 15-minute semi-structured interview, where participants discussed their experiences with the interfaces. These interviews also explored the factors that influenced their trust and included comparative evaluations of the interfaces, allowing participants to offer suggestions for improving the delivery of trustworthy health information. The interview protocol used in Study 2 is included in Appendix 5.C.1.

Data analysis

For the quantitative data, we first checked the data for normality using the Shapiro-Wilk test [208] and for homogeneity of variance with Bartlett's test [209]. We then performed a descriptive analysis to capture the essential data features. Subsequently, we utilized Pearson correlation analysis [311] to identify the significant relationships among key variables. To explore differences in trust levels across interfaces and types of search tasks, we conducted a Mixed Linear Model (MixedLM) regression [312] controlling the variables such as the usability level of the interface based on the correlations analysis. This approach provided a more detailed understanding of the factors influencing perceived trust in the information provided. Lastly, we conducted a mediation analysis to determine whether usability levels played a mediating role in influencing trust across interfaces.

For the qualitative data, we employed a thematic analysis [304] to examine interview responses, aiming to understand how trust in health information varies with different interfaces. To maintain consistency with Study 1, we followed the same analysis approach described in Section 5.3.1.

5.4.2 Quantitative Findings: Lab Sessions

Descriptive statistics

We conducted a descriptive analysis to examine the characteristics of this study sample. Detailed results are presented in Table 5.4.

Our findings showed that participants generally had a positive attitude towards technology, with an average PPT score of 3.87. eHealth literacy and AI literacy were also moderately high. Besides, usability ratings were highest for the text-based inter-

	Text-based Mean (SD)	Speech-based Mean (SD)	Embodied Mean (SD)
Propensity of trust in technology (PPT)		3.87 (.33)	
eHealth literacy		3.68 (.47)	
AI literacy		3.70 (.26)	
Prior experience (familiarity)	3.60 (.75)	2.35 (.88)	1.60 (.75)
Trust in information	4.19 (.42)	4.13 (.46)	4.00 (.48)
Trust in user interface	3.73 (.37)	3.57 (.46)	3.56 (.44)
Usability level of user interface	4.05 (.43)	3.55 (.67)	
Intention to use	3.55 (.86)	3.05 (1.07)	2.70 (.95)
Accuracy (speech recognition)	/	83.3%	88.3%
Follow-up questions (out of 60)	10	7	14

Table 5.4: Descriptive statistics of the measured variables in the lab study. The Propensity of trust (PPT), eHealth literacy and AI literacy were assessed during the pre-survey phase. (All measures above used a 5-point scale ranged from 1 to 5)

face, compared to the speech-based and the embodied interface, indicating that participants found the text-based interface easier to use, which correlated with their intention to use these interfaces for health information. Regarding the familiarity level with the interfaces, participants were most familiar with the text-based interface, followed by the speech-based and embodied interfaces. In terms of the intention to use the specific interface for health information search, participants most intended to use a text-based interface, followed by the speech-based and embodied interfaces.

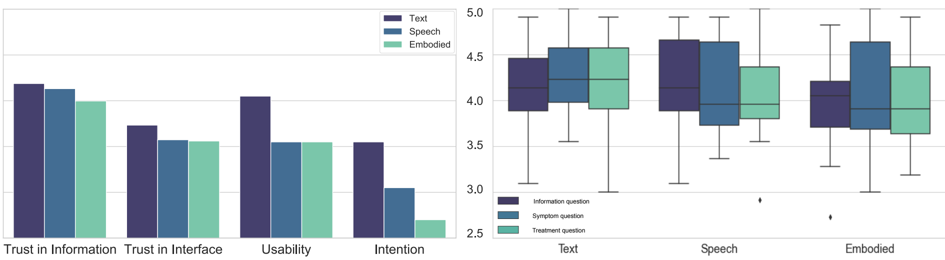


Figure 5.6: Left: Mean scores for perceived trust in health information, trust in the interface, usability, and intention to use, across three interfaces (text-based, speech-based, and embodied). Right: Median trust scores in health information across three search task types (information, symptom, and treatment-related) for each interface. Statistical comparisons across conditions are detailed in Table 5.5. Significant differences (**p<.01, *p<.05) were found between interfaces on trust, usability, and intention measures.

Figure 5.6 (left) summarizes mean trust levels in information and interfaces and usability levels. The right figure shows median trust levels in health information across three types of health-related search tasks: informational, symptom-related, and treatment-related information. Participants trusted the health information from the text-based interface the most, followed by the speech-based and embodied interfaces. Trust in the interfaces themselves showed a similar pattern: highest for the text-based interface, then the speech-based, and lastly the embodied interface. Trust levels were relatively consistent across different types of search tasks in all three interfaces as shown in Figure 5.6 (right), indicating similar trust in these types of search questions regardless of the interface used.

Additionally, we assessed the recognition accuracy of the speech-based and embodied interfaces, with each participant completing three tasks per interface, totaling 60 tasks each. For the speech-based interface, 10 tasks required one repetition and 3 required two, resulting in an accuracy of 83.3%. The embodied interface had 7 tasks needing one repetition and 3 needing two, with an accuracy of 88.3%. Follow-up queries were most common with the embodied interface (14 tasks), followed by the text-based (10 tasks) and speech-based interfaces (7 tasks).

5

Correlation analysis

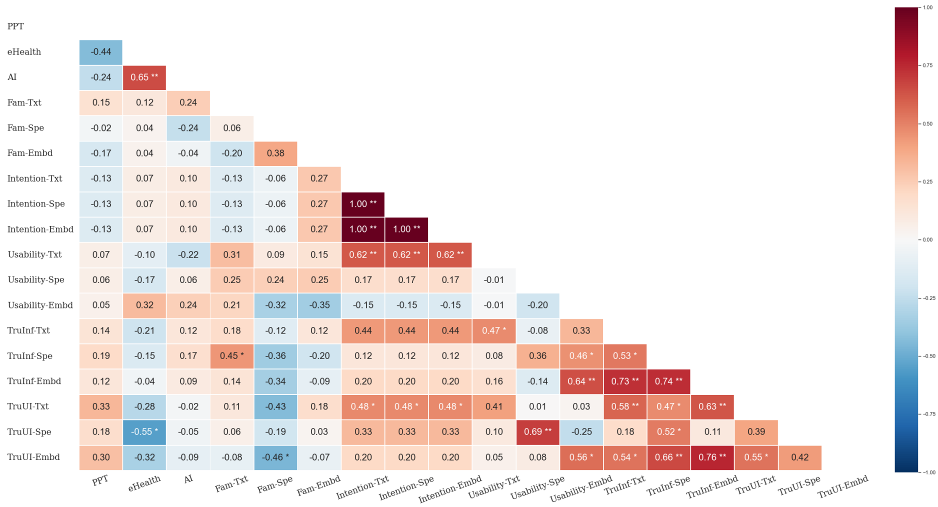


Figure 5.7: Pearson correlation with key variables (**p<.01, *p<.05). Note: "PPT" means participants' propensity of trust in technology; "eHealth" and "AI" are participants' eHealth and AI literacy; "TruInf" represents human perceived trust level in health information obtained; "TruUI" means human trust level in the user interface; "Fam-" is the familiarity level of each ; "Txt, Spe, Embd" reprent three CUIs respectively.

The results of Pearson correlation analysis in Figure 5.7 revealed relations among key variables.

For the text-based interface, there was a significant correlation between trust in the information and usability level ($r(20) = .47$, $p < .05$), as well as between trust in the information and trust in the interface itself ($r(20) = .58$, $p < .01$). In the case of speech-based interfaces, trust in the information was significantly correlated with trust in the interface ($r(20) = .52$, $p < .05$). Trust in the interface, in turn, showed a significant correlation with usability level (r(20) = .69, p < .01). For embodied interfaces, trust in the information significantly correlated with trust in the interface itself (r(20)=.76, p<.01) and usability level (r(20)=.64, p<.01). Additionally, trust in the interface was significantly linked to usability level (r(20)=.56, p<.05). These results highlight a consistent pattern: across all types of interfaces, trust in health information was closely related to trust in the interface. Additionally, trust in the health information was strongly connected to the usability level, particularly for text-based and embodied interfaces.

Human trust in information differs by dissemination interfaces

Scores	Coef. (β)	Std.Err.	P-value
Trust in information (Text vs. Embodied)	.189	.090	**.035**
Trust in information (Text vs. Speech)	.009	.072	.898
Trust in information (Speech vs. Embodied)	.180	.058	**.002**
Trust in interfaces	.213	.121	.078
Usability level	.187	.064	**.004**
Familiarity level	.064	.035	.062

Table 5.5: Results from the Mixed Linear Model (MixedLM) analysis. The coefficient ("Coef.") represents the estimated effect size of a predictor variable on the dependent variable ("trust in information"). The P-value represents the significance level of the differences.

The correlation analysis identified significant relationships between usability, trust in the interface, and trust in the information provided. Following these findings, we applied a Mixed Linear Model (MixedLM) [312] regression to control for variables such as usability, trust in the interface, and familiarity in each interface.

The MixedLM indicated significant differences of trust in health information disseminated between text-based and embodied interfaces ($\beta = .189, p = .035$) and between speech-based and embodied interfaces ($\beta = .180, p = .002$). These differences were largely influenced by usability which can significantly predict the trust in the information ($\beta = .187, p = .004$). There was no significant difference in trust levels across various types of search tasks ($p = .510$), suggesting that trust remains consistent regardless of the type of health-related question. A mediation analysis further confirmed that usability mediates the trust in information from various interfaces with indirect effect ($\beta = .154, p < .001$) and direct effect ($\beta = .098., p = .277$), suggesting the full mediation. These findings highlighted that usability and trust in the interface are crucial factors affecting trust in information.

5.4.3 Qualitative Findings: Semi-Structured Interviews

This section outlines themes from the interviews: First, we explored *"what factors impact trust perceptions in health information disseminated by different interfaces"*. Second, we explored *"how to improve LLM-powered CUIs for trustworthy health* *information seeking".*

What factor impacts trust perceptions in health information disseminated by different interfaces?

Prior experience and familiarity with the interface influence trust in information. Prior experience and familiarity with the interface are key factors in establishing trust in the information, as supported by our quantitative findings. Interviews indicated that participants' trust was primarily influenced by their familiarity with different interfaces and their habits in seeking online health information (P2-3, P5, P8-10, P11-12, P14, P16). For example, participants who frequently used Google or professional health websites found text-based interfaces more familiar and trustworthy. Moreover, the resemblance of text-based interfaces to widely-used social media platforms like WhatsApp and their similarity to consulting human health professionals or reading professional literature (P3, P6, P7) further increased trust in these interfaces.

> *"I think just about the presentations because the way I chat with the chatbot is more reliable because it provides me the information like I used to, uh, familiar with." - P2*

5

Usability of interfaces greatly impacts trust in the information. As indicated by our quantitative findings, particularly with text-based interfaces where usability mediated trust, participants favored text-based interfaces for their simplicity and ease of use (P1, P4-P7, P9-11, P18). While trust in text-based applications has already been established with the rise of LLMs, usability is crucial for building trust in less familiar speech and embodied interfaces (P1, P13, P18, P20).

> *"I think this is more likely to use it. [...], because more likely to use anything like add supplementary questions, you can sort of build trust on the information." - P18*

Information presentation style impacts the trust in information delivered by interfaces. The way information is presented significantly influences trust, especially in speech and embodied interfaces. Although the information source was consistent across all interfaces to prevent content bias, participants noted that while numbered formats work well for text, they may cause information overload in vocal formats, leading to cognitive overload and forgetfulness (P5-6, P10, P12, P20). Adjusting presentation styles, such as using storytelling or summarizing key points at the start of interactions can improve engagement and trust in these interfaces (P5, P10, P13).

> *"It is better to present quite an amount of information more like storytelling I would say, [...] Just like an essay like instead if you don't put it in bullet points but you say like you use connection words." - P10*

Information modality influences how information is processed, which in turn affects trust. The way people process different information modalities significantly affects their trust. Participants found it easier to process text than vocal information, as reading required less effort than listening (P5, P10, P12, P20). This ease of processing allows users to focus more on the content, which enhances trust, *I prefer to read book paper-wise rather than listening to it. - P12* Text-based information also mitigates issues related to context retrieval and timeliness found in speech, enabling users to cross-compare information within the same interface, thus deepening their understanding and building trust incrementally (P1, P4, P5, P7, P9, P11, P14, P16). Additionally, text-based interfaces allow easy sharing of information with health professionals or friends (P11) and facilitate detailed exploration and clarification of complex medical terms.

> *"If I don't even know what I missed, I don't know what kind of questions to follow up, whereas text one, I can see everything and I know exactly what I want to ask questions about. Because I can see and go back to and refer to the information if I forget anything." - P7*

Additionally, participants expressed that the embodied interface requires processing both verbal information and non-verbal cues like facial expressions (P3-P5, P7-P8, P11-12). This extra information processing can raise cognitive load [313], further affecting users' trust perception of the delivered health information.

> *"If I'm asking a question (to embodied and speech interfaces), I have to listen because I don't know what's coming next. [...], you had to actively listen. And if you missed one part, then you cannot go back." - P5*

Human-like features introduce additional considerations for trust. Participants reported that compared to text-based and speech-based interfaces, a physical embodiment gave a stronger sense of human interaction, leading to greater engagement and more follow-up questions (P3-4, P10). Our quantitative finding supports that participants asked more follow-up questions with the embodied interface than the other two. However, physical embodiment introduces factors that influence trust perception, such as the voice and appearance of the interface (P1, P3-P8, P11). Besides, the embodied interface can offer advantages similar to interacting with health professionals, such as observing symptoms and providing personalized advice, which text or speech interfaces lack. However, trust in human-like communication can vary. Some participants may distrust humans, leading to negative perceptions of human-like interactions. These benefits also present challenges to privacy concerns associated with its use and personal data leakage (P3, P5, P7, P11, P13-14). As P3 expressed worry about data sharing:

> *"Maybe it (embodied) can recognize my sound also track some information and like store the information in which may cause some like security problems maybe." - P3*

Overall, participants preferred embodied interfaces that balanced human-like and robotic features, favoring those that appeared neutral yet professional, with well-matched voices and movements.

> *"It is like the uncanny valley, I don't know how to say that, but there is the balance to bring, not be in the horror things but also (the interface) can be advanced." - P3*

How to improve LLM-powered CUIs for trustworthy health information seeking?

Enhancing features to improve information credibility. Participants were primarily concerned about the credibility of source information (P1, P4-5, P7, P9-10, P13-16), with distrust often stemming from a lack of references or the use of fake ones (P7, P9). They suggested adding features that link to original sources, allowing users to verify information, thereby increasing trust.

> *"I think if they can tell me this was taken from this XYZ source and cross reference with XYZX number of sources, I would probably trust it." - P7*

Additionally, CUIs can provide feedback that shows understanding of user queries, which can enhance trust in both the interface and the information provided (P5, P8). The CUIs can proactively ask contextually relevant follow-up questions (P2, P10) when they are not fully confident in understanding user's queries.

> *"It could be nice that you have like a light that's seeing when it's ready to get a response already. To get a question, and when it's talking, you have a different color or something so that you're assured that it gets your head has kept your question right or has captured your answer." - P8*

For embodied interfaces, participants recommended enhancing the interactive experience by synchronizing vocal interactions with more expressive features, as this could strengthen better relationship between the user and interface and further build trust (P1, P3-8, P11, P16).

Enriching interaction by integrating multiple information modalities and improving trust. Participants recommended incorporating multiple information modalities within a single interface. They felt that adding images or videos to text could improve comprehension, especially in health-related contexts (P4, P9). For instance, using images to show specific pain areas would enhance a text-based interface. In embodied interfaces, combining text, images, and videos was viewed as beneficial for boosting both trust and usability. This multi-modal approach could accommodate diverse searching styles and preferences, thereby enhancing the effectiveness of the health information delivered.

> *"It would be best I can do a little bit mixed with the text and the speech like type my questions and talk to the interface at the same time. [...] If there is a screen in front of the embodied robot to display the word is gonna give more trust feeling." - P4*

Personalize information toward individual needs. Echoing the qualitative findings from Study 1, participants in Study 2 also emphasized that health information should account for individual situations, such as personal symptoms or allergies, as well as regional differences, like variations in medical systems across countries (P2). They identified the generic nature of the information provided by interfaces as a barrier to trust and expressed a need for more personalized responses, particularly from embodied interfaces (P1-3, P7, P10-11). The trust would be significantly improved if interfaces could proactively inquire about personal symptoms to provide tailored advice (P2) and ask contextually relevant follow-up questions (P2, P10).

> *"I think my feeling of trust will increase because the information given to me is more personalized and it might be something that I can't even identify myself and therefore don't know to add into the speech or the text to search if the bot can see things that I cannot." - P7*

5.5 Discussion

This study explored how users' trust in health information is influenced by different search agents and dissemination interfaces. Through two mixed-methods studies, participants completed health information search tasks using two search agents (Google and ChatGPT) and various LLM-powered CUIs (text-based, speech-based, and embodied interfaces). We identified differences in trust levels across these interactions and gathered qualitative insights into the factors participants reported as influencing their trust ratings. In this section, we discuss the observed differences and their implications based on findings from both studies.

5.5.1 Exploring trust in health information from search agents to dissemination interfaces

In Study 1, participants showed on average a higher level of trust in ChatGPT compared to Google. Our findings showed that ChatGPT's conversational style and rapid, personalized responses made users feel heard and understood [248], which can be perceived as a sign of expertise and relevance [242, 314]. Additionally, the intuitive and user-friendly interface design of ChatGPT was reported to be easy to use and thus increases its credibility [119, 259]. As AI technology advances, such tools are increasingly seen as sources of cutting-edge information, influencing user trust [315]. Furthermore, we found a significant correlation between trust in health information provided by Google and trust in Google itself. This result might indicate a sign of traditional or longstanding trust in search engines like Google. Over the years, people have developed a strong relationship with Google and tend to view it not as actively interpreting or generating information, but simply as retrieving it. This perception contributes to sustained confidence in both the agent and the information it provides [269, 316]. In contrast, no such significant correlation was found for ChatGPT, suggesting that people might separate their trust in ChatGPT (or OpenAI more broadly) from their trust in the information it provides. This could mean that even though ChatGPT has advanced features, people are still taking time to develop a stable trust relationship with this new searching approach (e.g., newly released ChatGPT Search [317]) that shifts it closer to an active information provider. This distinction may also help explain why the Propensity to Trust Technology (PPT) had a greater influence on trust in ChatGPT than in Google. Users still develop familiarity with ChatGPT as an active information provider, whereas Google is more established and perceived more as a neutral retriever than an interpreter of information.

Study 2 aimed to establish the role of dissemination interfaces in shaping trust in health information. Participants expressed a strong preference for text-based interfaces, which interview findings revealed was due to familiarity and the ease with which the information could be verified. In contrast, while speech-based and embodied interfaces provided more natural interactions, they did not inspire the same level of trust, particularly in high-stakes health scenarios. Participants favored text-based interfaces for health inquiries because they allowed for more precise cross-referencing, a critical factor when seeking accurate medical advice. Quantitative analysis also supported these findings, showing that trust in an interface directly influenced trust in the information it provided. This findings aligned with previous research [254] highlighting that the credibility of a source, such as a brand, significantly impacts user trust in information. Similarly, studies [121, 125, 255] on web-based health information have demonstrated that the trustworthiness of the platform or source can heavily influence how users perceive the reliability of the information they receive. This highlighted the importance of designing LLM-powered tools that not only deliver accurate health information but also foster user trust with such tools.

Notably, we found no significant differences in trust levels across various health search tasks in both studies, indicating that trust in health information generalizes across specific query types. This finding was supported by the qualitative data, whereby participants reported similar trust levels across different types of health-related questions. These results highlight the importance of prioritizing the perceived reliability and credibility of search agents and interfaces in health contexts. However, participants demonstrated lower tolerance for errors in health-related scenarios compared to general contexts, underscoring the critical need for high reliability and accuracy in healthcare applications.

Although LLM-powered conversational search is still in its early stages, our participants already tended to trust its answers more than those from traditional search engines. Among various interfaces, the text-based format remained the most familiar and gained the greatest trust in information searching. As LLMs advance, understanding the reasons behind the higher trust level in such technology can provide practical insights for designing future LLM-powered health applications. Identifying key features of these agents and interfaces can further enhance trust in health information from such LLM-powered applications.

5.5.2 Role of search autonomy, prior knowledge, and experience in shaping trust of health information

Search autonomy is shaped by prior knowledge of health queries. The interview findings from Study 1 revealed that participants' preferences for search autonomy varied based on their prior knowledge, which suggested a close relationship between one's level of prior knowledge and the desire for autonomy in health information seeking. Participants appreciated Google's ability to cross-reference information from multiple sources, which allowed them to verify consistency and transparency, thus enhancing their confidence in the reliability of the information [287]. This feature was particularly valued by those who preferred a more autonomous approach to research, using Google to conduct in-depth comparisons. Conversely, participants noted that ChatGPT's ability to provide direct answers made it a more efficient choice, particularly for those with limited prior knowledge of search topics [318]. For these users, ChatGPT served as a convenient starting point, offering initial insights that guide further more detailed searches. This finding indicated that those with less background knowledge may value the simplicity and speed of LLM-powered search agents over the autonomy that Google provides. Additionally, prior knowledge may also serve as an internal calibrator to gauge trust, reducing the necessity for extensive autonomous searches. In the context of LLMs, follow-up questions to refine the information could be seen as a different form of search autonomy. These findings reflected individual differences in information-seeking behaviors and highlight that autonomy is crucial for those who prefer thorough research and verification, whereas prior knowledge can decrease the reliance on self-autonomous search.

Moreover, prior experience and familiarity level with search agents and dissemination interfaces are other well-established factors influencing trust perceptions, as supported by prior studies [254, 319–322]. The findings of Study 1 indicated that prior experience with traditional search engines like Google played a significant role in trust formation. As more people use ChatGPT, their familiarity with it will grow, and they may come to trust it increasingly. Similarly, quantitative results from Study 2 showed that familiarity with a user interface influences trust in the information it provides. Participants expressed greater comfort and trust in text-based interfaces because they aligned with their habitual methods of seeking health information. This prior experience could foster higher trust in text-based interfaces, while speech-based and embodied interfaces, despite their innovative features, still lag in gaining equivalent trust levels in health contexts. These findings emphasized the importance of interface design in establishing and maintaining trust. Prior experience with an interface can lead users to transfer their trust from the platform or tool itself to the information it provides. As such, building positive familiarity and aligning new technologies with users' existing search habits can be crucial in fostering trust in emerging conversational searches. As people use LLM-powered applications like ChatGPT more frequently for health information, their growing experience may influence how they search online. Studies suggest that users adapt their behavior depending on the tool; for example, they tend to ask more questions in text settings compared with voice assistants [323]. This shift in search agents presents a new research avenue to explore changing behaviors in using LLM-powered tools like ChatGPT for health information in daily life. Beyond experimental settings, a dynamic perspective is needed to understand how these tools are adopted, adapted, and integrated into the practices of health information searching.

5.5.3 Connecting usability and multi-modal processing to build trust in health information

Grounded in the work of modeling trust [254], usability strongly contributes to trust formation. Previous research has shown that [324] usability (e.g., ease of use) affects the feeling of trust in search agents. Usability emerged as a crucial factor in shaping trust in health information in our work, influencing both the effectiveness of search agents like Google and ChatGPT, as well as the CUIs through which information is disseminated. Findings from Study 1 suggested that usability directly impacted how users perceive the trustworthiness of health information. Participants appreciated Google' s usability for its ability to cross-reference sources, providing a transparent and thorough search process, while ChatGPT's streamlined approach was preferred by those seeking convenience, particularly for quick answers or initial insights. The study by Xu et al. [318] supports this, showing that users spend less time on ChatGPT for similar tasks compared to Google, indicating a preference for its efficiency. Thus, when usability is defined by simplicity and speed, ChatGPT builds trust by offering a more direct path to information.

Beyond the search agents, usability also plays a pivotal role in how CUIs manage different modalities of information, as evident in Study 2. Multi-modal CUIs, such as speech-based and embodied systems, represent a shift from traditional text-based interfaces, introducing a richer yet more complex interaction experience. Prior studies highlight that multi-modal interfaces can enhance engagement by integrating visual and auditory elements next to text alone [325–328]. Deldjoo et al. [327] indicate that multi-modal systems can facilitate more comprehensive and immersive comprehension of content by stimulating multiple senses. Robb et al. [313] also explore that the multi-modal embodied agents can offer more engaging interaction thereby increasing perceived trustworthiness compared to the speech-based interfaces. However, this complexity comes with increased cognitive load [296, 313], as users must process and interpret multiple forms of input as explored in prior work [328, 329]. It could influence the perception and trust of the information which aligns with the MAIN model from [330]. Our findings from Study 2 supported that the added complexity can impact usability and subsequently, trust perception. While multi-modal interactions may feel more dynamic and human-like, they can also introduce opportunities for errors, such as misinterpreting tone in speech [293, 294, 331, 332] or synchronization issues in embodied interactions [333–337]. In contrast, text-based interfaces offered a simpler, more predictable experience. As noted by participants in Study 2, the straightforward nature of text reduces the risk of misunderstandings, fostering greater transparency and reliability in information delivery. Moreover, consistency in information and presentation style can also affect trust perception as indicated in [296]. Any perceived mismatch or inconsistency can lead to skepticism or distrust, a challenge generally absent in text-based communication. Therefore, designers should prioritize the usability design to foster trust in digital tools for health information search.

5.5.4 Anthropomorphic intelligence of LLMs for health information seeking

Human-like features

The rise of LLMs has sparked discussions around the anthropomorphism of these technologies [250, 338–341]. Research shows that human-like qualities in LLM-powered agents, such as ChatGPT, can significantly influence users' trust levels [168, 284, 342–344]. A key anthropomorphic feature of ChatGPT is its conversational interaction style, which helps create intuitive information delivery and builds a trust-based relationship between users and the agent. Additionally, role-playing abilities [345, 346], such as assuming the persona of a doctor, can make interactions feel like a real consultation, enhancing perceived relevance and trust in the information provided [347]. Replicating human conversational patterns can further enhance trust. Participants of Study 1 indicated that LLMs should communicate in a human-like manner by breaking down complex information into manageable pieces rather than overwhelming users with lengthy responses. Cai et al. [348] found that mirroring the interaction style of human advisors can increase user attention and comfort, a point that our participants echoed. Moreover, incorporating empathy, encouragement, and understanding in AI interactions can further foster deeper user engagement and trust [349–351].

However, not all research supports full anthropomorphism as a requirement for trust. Placani [251] and Festerling et al. [352] suggest that using familiar language, a reassuring tone, and a patient-centered approach, similar to traditional healthcare, can be more effective for fostering trust [320, 321]. Troshani et al. [353] and Placani [251] similarly argue that making LLM-powered AI highly human-like in appearance does not necessarily enhance trust, particularly in health contexts where users often prefer neutral, clear information [33]. Human-like features, such as facial expressions, can evoke discomfort or skepticism due to the "uncanny valley" effect [249], where minor mismatches between appearance and behavior can undermine trust. Participants in Study 2 emphasized the importance of transparency, preferring clear identification of AI as the source of information. Knowing that the information was generated by an LLM helped participants set realistic expectations and critically assess the content. This findings aligned with previous research highlighting the importance of source identification for trust [245, 251, 255]. While anthropomorphic features can make interactions feel more relatable, they also risk blurring the distinction between human and machine, potentially leading to confusion about the origin of information. Ensuring that users know they are interacting with AI helps establish trust by fostering informed expectations. Overall, these findings highlighted the need for a balanced approach that leverages AI's capabilities without overstepping into areas where human expertise is crucial [33, 321]. Striking this balance can help maximize the benefits of AI interactions while maintaining transparency and trustworthiness in health information searches.

Personalized interactions

LLMs' ability to provide personalized interactions is another key to build user trust and engagement in health information searches. As prior research [24] suggests, people are more likely to trust personalized responses when seeking health information online. Our findings aligned with this, indicating that personalization indeed enhances trust. Specifically, our study showed that personalization can be achieved through various elements, including language style, interactive dialogue that enabled follow-up questions, and a tone adapted to the user's emotional state or specific symptoms. Personalized language style in LLMs can extend to personal and cultural sensitivities, fostering a more inclusive interaction [345, 346, 354]. Unlike traditional search engines like Google, LLMs interpret complex queries with contextual depth, allowing for responses that are tailored to individual needs. This deeper understanding enhances users' sense of being understood, thereby increasing trust. Additionally, LLMs handle follow-up questions effectively, maintaining continuity in interactions and improving user experience [286]. This is particularly valuable in health contexts, where advice tailored to specific symptoms or emotions can make LLM-powered agents or tools appear more empathetic and supportive.

Despite these benefits, LLM-powered conversational searches present significant privacy challenges [297, 355]. Our interviews from Study 2 revealed that human-like features raise additional concerns, particularly around privacy, with participants expressing worries about potential personal data leaks. Personalized and multi-modal CUIs, which often collect sensitive data, further amplify privacy risks. Features like voice recognition and facial analysis [356, 357] necessitate transparent data collection practices and explicit user consent, especially in health contexts. While text-based interfaces provide more anonymity, speech and embodied interfaces inherently gather more personal data, increasing the potential for misuse or breaches. As Bansal et al. [267] note, trust in these interfaces depends not only on content credibility but also on how data is managed. Any perceived misuse or lack of security can severely undermine trust. Participants in Study 2 echoed concerns about constant listening in speech-based interfaces and the potential for visual monitoring by embodied interfaces, particularly when these actions occur without explicit consent [358]. Given that personalized LLM-powered tools collect a wide range of sensitive data, from text and speech to biometrics, robust data protection and transparent privacy policies are essential [37, 287, 329, 359]. Addressing these concerns is critical to fostering trust and ensuring that the benefits of personalized interactions do not compromise ethical transparency and privacy.

5.6 Limitations and future work

While our study provides valuable insights into the difference in trust in online health information from distinct search agents and across interfaces, it is important to acknowledge its limitations for a more nuanced understanding of trust.

First, our research focused solely on Google and ChatGPT as representatives of traditional search engines and LLM-based conversational search, respectively. While these are prominent examples, it is important to note that there are hybrid approaches emerging in the field, like those seen with Bing [360] and the newly released ChatGPT Search [317], combining web browsing and LLMs to offer an enhanced search experience. Nevertheless, our study provides a crucial first step toward unraveling insights into human trust and reliance on information provided by traditional and LLM-powered search agents (even if separately), paving the way for future research to explore emergent hybrid LLM search technologies and their implications on human trust.

Second, our study focuses on health information, which is an important yet specific domain. The differences in trust perception may vary when considering other types of information or contexts, and this specificity could limit the generalizability of our findings.

Third, our study does not evaluate the accuracy of responses from LLMs. LLMs may make up answers (termed as AI "hallucination" phenomena [361]), which is one concern shared by our participants. This invites future work into trust dynamics across other information-seeking contexts, to examine more closely the accuracy of LLM responses and their impact on trust. The web searching capability in the latest version of LLMs such as ChatGPT Search [317] might be able to offset some of these concerns, though they remain fallible.

Fourth, our research design involved a lab study followed by semi-structured interviews. While this method allowed for in-depth insights, it may not fully capture the complexity and spontaneity of real-world information-seeking behaviors. While it would be ideal for participants to engage in naturalistic search behaviors based on their needs, such approaches raise significant privacy concerns. To mitigate this, our study design is purposefully structured to not only address privacy concerns but also offer a controlled environment to understand trust development in online health information seeking.

Lastly, the influence of varying demographics, like age and education level, was not fully considered. Although these factors could potentially influence trust in online information, addressing these variables was beyond the scope of our current work. Nevertheless, it provides an opportunity for subsequent research to explore how different demographic groups perceive and trust online health information differently, thus broadening the understanding of online health information trust across diverse user groups.

With the continuous and rapid advancements in AI technology, it is vital to continue validating and verifying such findings relating to trust perceptions, and how these may change over time with newer AI advancements.

5.7 Conclusion

In this research, we investigated trust in health-related information by comparing two search agents (Google and ChatGPT) and three LLM-powered conversational user interfaces (text-based, speech-based, and embodied interfaces). As online and conversational searches become critical for health information seeking, understanding trust perceptions is vital. Our mixed-methods Study 1 revealed that participants had greater trust in ChatGPT over Google, influenced by factors such as prior experience, information presentation style, and interaction mode, while the type of health query did not significantly impact trust. In Study 2, we also examined how different user interfaces affect trust in health information dissemination. Text-based interfaces were most trusted due to their familiarity level and ease of use, while speech-based interfaces offered convenient and natural verbal interaction. Embodied interfaces provided a more immersive experience but raised concerns about privacy and authenticity. Overall, our findings emphasize the importance of both source agents and dissemination interfaces in shaping trust. As generative AI and LLM-powered tools evolve, designing health tools that are reliable, user-centered, and mindful of trust dynamics will be crucial for their successful adoption. Our work provides a foundation for future research to further refine digital health applications, ensuring they are not only reliable and informative but also trusted and tailored to diverse user needs effectively.

Chapter appendix

5.A Search Task

General Questions	Symptom-Cause-related Questions	Treatment-related Questions
What is (are) Low Blood Pressure ?	What are the symptoms of Heart failure - overview ?	What are the treatments for Heart failure?
Do you have information about Vitamin D	What are the symptoms of Type 2 diabetes ?	What are the treatments for Type 2 diabetes ?
Do you have information about Weight Control	What are the symptoms of Dry eye syndrome ?	What are the treatments for Coronary heart disease ?
What is (are) Chronic Pain ?	What are the symptoms of Coronary heart disease ?	What are the treatments for Persistent depressive disorder ?
What is (are) Migraine ?	What are the symptoms of Frozen shoulder ?	What are the treatments for Frozen shoulder ?
What is (are) Coronavirus Infections ?	What are the symptoms of Separation anxiety in children ?	What are the treatments for Chronic fatigue syndrome ?
Do you have information about Quitting Smoking	What are the symptoms of Natural short sleeper ?	What are the treatments for Obsessive-compulsive disorder ?
What is (are) Hair Loss ?	What causes Memory loss ?	What to do for Memory loss ?
What is (are) Fever ?	What are the symptoms of Burns ?	What to do for Toothaches ?
What is (are) Mental Disorders ?	What causes Obesity ?	What to do for Sweating ?
What is (are) Food Allergy ?	What are the symptoms of Dislocation ?	What to do for Hair loss ?
What is (are) Sleep Disorders ?	What are the symptoms of Allergies ?	What are the treatments for Dementia ?
Do you have information about Diabetic Diet	What are the symptoms of Heart block ?	What are the treatments for Obesity ?
What is (are) Stress ?	What causes Toothaches ?	How to prevent Dislocation ?
What is (are) Diabetes ?	What causes Sweating ?	What are the treatments for Common cold ?
Do you have information about Exercise and Physical Fitnes	What causes Hair loss ?	What to do for Fatigue ?
What is (are) Depression ?	What are the symptoms of Sprains ?	What to do for Fever ?
What is (are) Heart Diseases ?	What are the symptoms of Sleepwalking ?	What are the treatments for Flu ?
What is (are) Obesity ?	What causes Type 2 diabetes ?	How to prevent Miscarriage ?
What is (are) Dislocations ?	What causes Coronary heart disease ?	What are the treatments for Food allergy ?
What is (are) Flu ?	What are the symptoms of Persistent depressive disorder ?	What to do for Overweight ?
What is (are) Fatigue ?	What are the symptoms of Personality disorders ?	What are the treatments for Sleepwalking ?
What is (are) Headache ?	What are the symptoms of Chronic fatigue syndrome ?	What to do for Sprains ?
Do you have information about Cardiac rehabilitation	What are the symptoms of Dementia ?	What are the treatments for Cat-scratch disease ?
Do you have information about Pregnancy and Medicines	What causes Weakness ?	What to do for Burns ?

Figure 5.A.1: Health-related questions used as the search tasks in Study 1 and 2.

5.B The Interview Protocol for Study 1

Study 1: Trust in Search Agent - Interview Protocol

Introduction

"Thank you for taking the time to participate in this interview. I'm XXX, the researcher working on this study about trust in online health information. Today we'll talk about how you generally search health information online with different search agents"

"Our conversation will take approximately 25 mins and you are free to interrupt or ask any questions at any time. There are no right or wrong answers, we encourage you to share your honest opinions."

Warm-Up Question

1. To begin the interview, could you first describe how you typically search for health-related information online?"
    - What tools or applications did you usually use?
    - What is your general feeling when you use such tools for health information seeking?

Main Interview Questions (Factors Affecting Trust)

Google search:

2. Let's be more specific on your search process with Google in this study. What is your experience or feelings of using Google for health information seeking? (Negative/Positive)
    - How hard is it to find the answer? Why?
    - How do you feel about the way of interaction with Google for health information seeking?
3. How do you feel about the information you got from Google?
    - Are you satisfied with the answers?
    - Do you think they are trustworthy? Reliable?
        - What factors can increase or decrease your trust in the information found on Google?

ChatGPT search:

4. Did you use tools like ChatGPT before in your daily life to search for health information?
    - If yes:
        - How do you use that to search for health information?"

1

Figure 5.B.1: The Interview Protocol for Study 1.

5.C The Interview Protocol for Study 2

Study 2: Trust in Dissemination Interface - Interview Protocol

---

Introduction

"Thank you for taking the time to participate in this interview. I'm XXX, the researcher working on this study about trust in digital health information. Today we'll talk about how you generally search health information through different user interfaces"

"Our conversation will take approximately 20 mins and you are free to interrupt or ask any questions at any time. There are no right or wrong answers, we encourage you to share your honest opinions."

Warm-Up Question

1. Can you describe your initial feelings about receiving health information from each type of interface (text-based, speech-based, embodied interfaces)?
    - Which interface do you find more intuitive or comfortable for seeking health information, and why?

Comparison of Trust in Interfaces

2. Among the three interfaces, do you trust one interface more than the others for obtaining health information? Why?
    - Do you think some interfaces seem more trustworthy or less?
    - What are the differences in feelings of trust you felt when you searched for health information by different interfaces? And why?
3. Do you think the usability affected your trust in the health information provided?

Trust in health information:

4. Did you find the information provided by one type of interface more trustworthy than the others?
    - Did the modes of information delivery (text, voice, physical presence) affect your trust in the health information provided?
5. What specific elements or features of each user interface do you feel influenced your level of trust in the health information provided?
6. Do you think the physical embodiment of a bot can influence your trust perception of the information? Why or why not?
    - How does the lack of physical presence in text and speech-based interfaces affect your trust in the information they provide?

1

Figure 5.C.1: The Interview Protocol for Study 2.

6

Understanding Trust toward Human versus AI-generated Health Information through Behavioral and Physiological Sensing

This chapter is based on the following publication:

Authors: Xin Sun, Rongjun Ma, Shu Wei, Pablo Cesar, Jos A. Bosch, Abdallah El Ali.

Original title: Understanding Trust toward Human versus AI-generated Health Information through Behavioral and Physiological Sensing

Submitted to: International Journal of Human-Computer Studies (Major revision)

6

Abstract

As AI-generated health information proliferates online and becomes increasingly indistinguishable from human-sourced information, it becomes critical to understand how people trust and label such content, especially should such information be inaccurate. We conducted two complementary studies: a 2 (source: Human vs. LLM) × 2 (label: Human vs. AI) × 3 (type: General, Symptom, Treatment) mixed-methods survey (N=142) and a within-subjects lab study (N=40) incorporating eye-tracking and physiological sensing (ECG, EDA, skin temperature). Participants viewed health information with varied source-label combinations and rated their trust, while their gaze behavior and physiological signals were recorded. We found that LLM-generated information is trusted more than human-generated content, whereas human labels are trusted more than AI labels. Trust remained consistent across information types. Eye-tracking and physiological responses varied significantly by source and label. Machine learning models trained on these behavioral and physiological features predicted binary self-reported trust levels with 73% accuracy and information source with 65% accuracy. These findings demonstrated that adding transparency labels to online health information modulates trust. Behavioral and physiological features showed potential to verify trust perceptions and indicate if additional transparency is needed.

6.1 Introduction

The internet has become a primary source of health information [23, 24], with 58.5% of American adults [25] (survey on 2022) and 55% of Europeans [26] (survey on 2022) using online sources for health-related searches. This shift has transformed how individuals access and engage with health-related content. The rise of digital platforms, including professional medical websites [236, 273] and AI-driven tools like health chatbots powered by Large Language Models (LLMs) [1], has made health information more accessible and convenient than ever. As such, understanding how different information sources affect trust perceptions is increasingly critical [48], as the source is a key factor in determining trust in health contexts [47, 256], where such information can significantly impact health-related decisions involving inherent risks [37, 38]. Besides, misleading labels or unclear sourcing may result in misinformation and poor health decisions [28, 38], even if AI system disclosures are increasingly mandated by regulations, such as the European AI Act [27]. Research shows that trust perceptions differ between human-generated and machine-generated content across individuals [362–364], and that disclosed labeling (e.g., with/without indicating AI involvement) can greatly impact trust [365]. While some studies suggest that people may prefer algorithmic or AI over human judgments [366], other studies show that AI labels can decrease trust [362, 367, 368].

Despite growing interest in AI-generated health content, existing research lacks empirical insights into how information sources and labeling interact to influence trust in the health context, particularly in the era of LLMs. Understanding how both the source and label influence trust is crucial, especially for personal health decisions that directly impact individuals' health and well-being [38]. By examining explicit self-reports, this work aims to clarify how these factors jointly influence people's perceived trust and to inform the design of AI-powered health information systems that promote responsible usage through transparency labels [369].

In addition, human trust is inherently subjective. Although self-reported measures are widely used due to their simplicity and directness, research by Chen et al. [53] and Kohn et al. [54] highlight that self-reported trust measures inherently are vulnerable to reporting biases like social desirability bias and Initial Elevation phenomenon [55]. These biases can compromise the reliability and validity of self-reported trust assessments. In comparison, behavioral and physiological measures may provide an alternative and perhaps more objective lens for understanding trust perception [56–58]. Use of this modality also aligns with recent interest in Human-Computer Interaction to draw on physiological sensing when designing or evaluating interactive systems [59]. Behavioral patterns such as eye movements

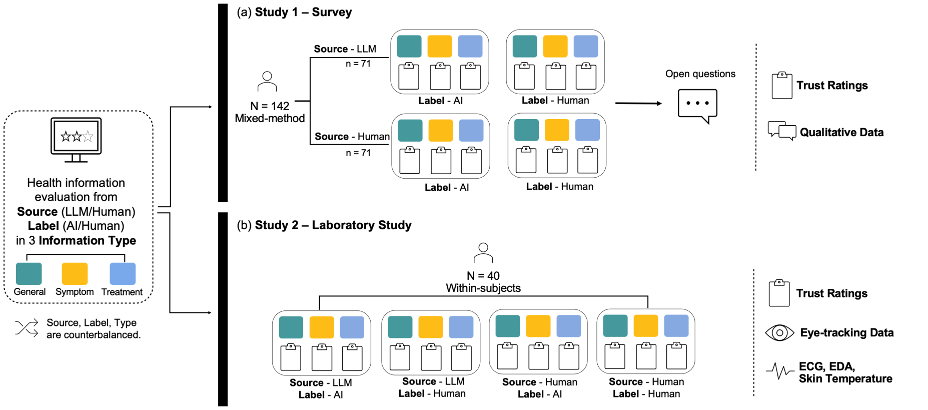


Figure 6.1: Visual summary of the studies in this paper. (a) Study 1: Mixed-methods crowdsourcing survey study to measure perceived trust; (b) Study 2: Within-subjects lab study to measure perceived trust, as well as behavioral and physiological responses.

and physiological responses, as assessed by Electrocardiogram (ECG) [56] and Electrodermal Activity (EDA) [62], could reveal how individuals process information and make trust-related decisions in health contexts. For example, eye movements, such as fixation duration and saccade behaviors, can indicate cognitive engagement with the information, while physiological responses like heart rate variability (HRV) [56, 60, 61] and skin conductance levels (SCL) can reveal emotional arousal and stress responses. These outcome measures may further help interpret user trust perceptions [61, 62]. Thus, exploring these behavioral and physiological indicators can contribute to a more comprehensive understanding of trust formation in digital health contexts [56, 57, 370] and further, help develop strategies to enhance the trustworthiness of online health information, especially given the growing use of LLM-powered tools for health advice [239–241].

In this work, we adopted a mixed-methods approach (as shown in Fig 6.1) to investigate how the source and disclosed transparency (i.e., disclosed label) of personal health information affects trust perceptions. We thereby ask: ***(RQ1) Does the actual source, disclosed label, and type of personal health information influence people's perceived trust in online health information?*** To examine this question, we conducted an online crowdsourcing survey (N=142) using a 2x2x3 factorial design. Source (Human Professional vs. LLM) was treated as a between-subjects factor to minimize potential biases from participants directly comparing human and AI sources. In contrast, Label (Human Professional vs. AI) and health-information Type (General vs. Symptom- vs. Treatment-related) were within-subjects factors to enable a nuanced comparison of trust perceptions across different labeling and information types within the same participant. This mixed design balanced the reduction of cross-condition biases with the sensitivity of within-subject comparisons. Participants rated their perceived trust in the health information they received using standardized self-report scales, which served as our primary outcome measure. Based on the key finding that both information source and label influence trust perceptions, we then ask: ***(RQ2) Can behavioral and physiological signals be used to understand trust perceptions toward human and AI-generated health information?*** Following a similar 2x2x3 design with fully within-subjects factors, we conducted a laboratory study (N=40) employing eye tracking (i.e., gaze patterns and pupil dilation) and physiological signals (i.e., based on ECG, EDA, and skin temperature), to understand the relationship among these indicators and trust perceptions in health information manipulated by source and label. By allowing each participant to serve as their own control, this design minimized variability due to individual differences and maximized the robustness of condition-specific inferences. Importantly, participants were not informed that labels could be intentionally mismatched with the actual source (i.e., cross-labeled) in both studies. This ensured that participants evaluated the health information and its disclosed label as presented, without being influenced by a heightened awareness of potential labeling errors, thereby allowing us to more accurately assess their trust perceptions on both information itself and its labeling.

Online survey findings showed that the (actual) source of information significantly influenced trust perceptions, with participants displaying higher trust in LLM-generated health information compared with human professionals. Second, the labeling of the source played a crucial role: health information labeled as coming from human professionals led to significantly higher trust than information labeled as from AI, i.e., regardless of the actual source. Third, the type of health question did not significantly affect trust, alone or in interaction with label and source. Together, these observations suggested that perceived trust is not influenced by the nature of the health query, and that the source and labeling of the health information are main determinants. The laboratory study supported the survey findings, with additional insights: gaze data, such as fixation count and duration, saccade count, and pupil diameter, varied significantly based on the source and labeling of health information. Moreover, physiological features, such as heart rate variability (HRV, measured as the root mean square of successive differences, RMSSD) and skin temperature, differed when participants engaged with information with different labels. These findings indicated that the source and labeling of health information influence both behavioral and physiological responses. Further prediction tasks were performed based on behavioral and physiological data, yielding 0.35 R^2 for predicting trust scores and 73% accuracy in classifying binary trust levels (high vs. low). Additionally, we achieved 65% accuracy in classifying the source of health information. These results underscored the potential of leveraging behavioral and physiological signals as an additional means to model and understand trust perception toward human vs. AI-generated health information.

Our exploratory work offers two primary contributions: **(1)** We provided empirical evidence showing that trust in online health information is influenced both by its actual source and disclosed label. **(2)** We found that trust perceptions in personal health information vary at behavioral and physiological levels, offering complementary insights beyond self-reported trust and helping to identify discrepancies between the explicit (i.e., self-reported) and implicit trust responses. To our knowledge, this is one of the few studies that combine physiological (e.g., HRV, skin temperature) and behavioral (e.g., gaze) signals to understand trust in AI-generated health information. Our work highlights the importance of considering AI transparency labels when measuring trust in health information and the vulnerability of trust abuse to mislabeling. It further opens the possibility of verifying trust perceptions and inferring if and when to apply transparency labels based on sensed behavioral and physiological data.

6.2 Related Work

6.2.1 Trust in Online Health Information Seeking

Trust is a multifaceted psychological construct essential to both interpersonal and human-technology interactions. Mayer, Davis, and Schoorman's integrative model of trust [41] defines trust as a willingness to be vulnerable to the actions of another party, based on the expectation that the party possesses the ability (competence), intends to do good (benevolence), and adheres to a set of principles that the trustor finds acceptable (integrity). In the digital age, trust extends to automation systems and technology in general. Work by Lee and See [43, p. 51] defines such trust as: "An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability". Similarly, McKnight et al. [252, 253] found that trust in technology is influenced by various factors such as perceived usefulness, ease of use, reliability, credibility, and risk. In the context of health, trust is particularly important due to the sensitive nature of health information and its impact on health-related decision-making, which can have dire health consequences should it be incorrect [37, 38]. Trust formation in these contexts is likewise complex and influenced by both intrinsic and extrinsic factors, including individual characteristics such as prior knowledge, health literacy, and decision-making roles, as well as external cues used to validate information, as highlighted by Vereschak et al. [371].

For instance, amounts of work [46, 48, 244, 256–258] indicated that the credibility of the information source is crucial, the design [119, 121, 125, 259] and usability [260] of the health-related tools can significantly affect trust. A well-designed, easy-to-navigate website or app is usually perceived as more trustworthy than a platform with a confusing or complicated user interface. User prior experience like familiarity levels [261], and user expectations [262] influence trust perceptions as well. Lastly, transparency, data privacy, and ethical considerations increasingly influence trust levels as well [170, 267–269, 372]. This body of work has been instrumental in systematically understanding how people develop trust in health contexts.

As a step towards bringing together current knowledge on people's perceived trust in online health information, Johnson et al. [254] indicates that trust in online health information is evaluated based on three main factors: source credibility, the context of information delivery (design and navigability of the health system), and the content itself (credibility, reliability, relevance). These components collectively influence the user's overall trust in health information. Building on this, subsequent research has delved deeper into the intrinsic quality of information, highlighting how factors such as perceived credibility [121, 125, 141, 143, 264], clarity of presentation [254], and relevance to users' health query [257, 261, 262, 265] further influence trust. These content-specific dimensions complement broader contextual and source-related factors to form a more complete understanding of trust in health communication [259, 263, 279]

Integrating the literature and as a step toward transparent AI communication, Liao and Sundar [359] introduced a conceptual model called MATCH that describes how trustworthiness can be communicated in AI systems through trustworthiness cues. They proposed a checklist that can aid technology creators in identifying reliable cues for inclusion, such as transparency and interaction design. Although this thesis does not directly apply MATCH, its emphasis on designing trustworthy cues that are both warranted and user-appropriate aligns with our purpose to explore how trust is formed and shifted between human and AI-generated or labeled health information.

6.2.2 Source and Label Transparency in the Age of LLMs

The internet has become a vital resource for health information [23], with websites like WebMD [373] and Mayo Clinic [236] providing expert-curated content. The rise of LLMs like ChatGPT [65] has revolutionized access to online health information by offering conversational interactions to health queries [29]. Trust in these LLM-powered tools is influenced by various factors [248], including the perceived credibility of their responses, the clarity of information, transparency about how the information is generated [27], and users' familiarity and experiences using such AI technologies [374]. Among these, information source (e.g., human-authored vs. AI-generated) plays a critical role in shaping trust. Research [46,48,375] have shown that trust is significantly affected by the perceived credibility of the information source. While LLMs have been effective in providing health information [343,374], concerns remain about their credibility and reliability. Although human professionals are traditionally viewed as authoritative and trustworthy due to their expertise [376,377], studies like Logg et al. [366] show that users may trust AI for specific tasks, and Shruthi et al. [368] indicated that people overtrust AI-generated medical responses. However, other research [365, 377] highlighted people's preferences for human-generated health advice, suggesting that trust varies based on context. Additionally, Montag et al. [378] found that trust in humans and AI may not be directly associated, suggesting people have distinct trust mechanisms for each. These varied trust levels underscore the complexity of trust formation towards the information from human and AI sources.

Labeling of information sources plays an additional key factor in shaping trust perceptions in the era of LLMs. Jakesch et al. [379] demonstrated that users perceive content as less trustworthy when it is labeled as AI-generated, even when the content quality is identical, which indicates that labeling influences how users perceive trustworthiness. Similarly, Reis et al. [365] found that perceived AI involvement significantly impacts trust in digital medical advice, as participants in their study were less willing to follow health advice when they believed it was generated by AI rather than a human expert. Studies by Walker et al. [362] and Kerstan et al. [377] have also shown that people tend to trust advice more when it comes from human professionals rather than from LLMs, especially when the source is explicitly stated. Yin et al. [367] found that while AI can create a sense of being heard, labeling content as AI-generated can reduce its perceived impact. These findings underscore how labels can significantly impact trust, even when AI performs tasks effectively. Furthermore, Scharowski et al. [380] explores the potential for AI certification labels (e.g., "Digital Trust Label" by the 2023 Swiss Digital Initiative), and finds that such labels can mitigate data-related concerns surfaced by end-users such as data protection and privacy, however this came at the cost of other concerns such as model performance, which poses it own challenges. Nevertheless, these works highlight that transparent communication about how AI systems operate and the data sources they use can further enhance or maintain trust among users [50,366].

As AI becomes more integral to health contexts, this work specifically examined the influence of source and labeling on trust in health information, offering insights for designing trustworthy AI-powered health information systems.

6.2.3 Behavioral and Physiological Signals for Understanding Trust Perception

Traditional research on trust perception has heavily relied on self-reported assessments; however, many studies [53, 54] suggest behavioral and physiological signals may add a relevant layer of information. Integrating these implicit measures helps offer a complementary understanding of trust in human and LLM-generated health information. For example, research by Kenneth et al. [66] shows that eye movement metrics like fixation, saccade, and pupil dilation provide insights into cognitive load and attention allocation during information processing. While these physiological indicators do not directly measure trust, they may reflect how users cognitively engage with content they perceive as more important, credible, or challenging. For instance, increased pupil dilation, linked to higher cognitive load [381] and emotional arousal, may suggest deeper cognitive processing, which may co-occur when individuals are evaluating information for trustworthiness or making health-related decisions. Although the relationship between trust and cognitive load is complex, monitoring these signals may help identify moments of increased scrutiny or hesitation, offering indirect cues about trust-related states. As an example of such research, Ji et al. [382, 383] demonstrated that physiological signals, such as electrodermal activity, blood volume pulse, and gaze, vary meaningfully across different information processing activities (e.g., reading, speaking, listening) during information-seeking tasks. Moreover, prior work has used behavioral data to explore how people engage with online news content, particularly in the context of misinformation. For instance, Abdrabou et al. [384] found that gaze and mouse movement patterns could help distinguish between user exposure to real versus fake news, achieving moderate accuracy in identifying subconscious engagement with misinformation. Similarly, Sumer et al. [385] showed that eye-tracking data reflected differences in how users read and process true versus false news articles, suggesting that such behavioral signals can offer objective indicators of how people implicitly respond to varying degrees of information credibility. Studies [54, 370, 386] demonstrate that distinct gaze patterns are linked to trust levels, with higher fixation counts and longer duration typically indicating focused attention, greater cognitive engagement and trust in the information [66, 370, 387, 388]. Saccades, characterized by the frequency and length of eye movements between fixations, often signal information verification processes [386, 389, 390]. These findings suggest that these multimodal implicit signals can be sensitive indicators of user engagement and cognitive effort, offering potential to infer user states such as trust or uncertainty in information processing contexts. Furthermore, using these signals can help prevent misinformation by identifying when users hold a lower trust level on information. For example, based on such eye movement data, systems can prompt users to reconsider or verify information before sharing. While this may not guarantee the reliability of the information itself, it introduces an additional layer of scrutiny that can help reduce the spread of misinformation and support more thoughtful decision-making.

Physiological features such as ECG [56], EDA [62], and skin temperature [61] can likewise be useful for understanding what are otherwise only self-reported trust perceptions. Heart Rate Variability (HRV), derived from ECG, reflects the level of arousal, with higher HRV indicating lower physiological arousal associated with relaxation, comfort, and higher trust levels [60, 391, 392]. EDA measures, including Skin Conductance Level (SCL) and Skin Conductance Response (SCR) are similarly tied to emotional arousal, where lower conductance is used to infer greater comfort and trust [61, 62, 370]. Similarly, changes in skin temperature are thought to reflect engagement levels, with higher temperature suggesting increased cognitive engagement with information [61]. As investigated by prior work, trust perception, a complex, subjective cognitive and emotional process, can be (objectively) assessed using machine learning models by analyzing physiological (e.g., ECG and EDA [393], EEG [57]) and behavioral (e.g., gaze patterns [58, 394]) indicators. These models help reduce subjective bias and can provide real-time insights into trust responses, least of which is an additional verification means alongside self-reports. For our work, we explore how trust is influenced by the source and labeling, and whether such signals meaningfully vary and in what capacity they can aid in predicting trust perceptions.

6

6.3 Study 1: Online Survey

6.3.1 Study Methods

Design

We conducted an online survey using a mixed 2 (IV1 - Actual Source: Human professionals vs. LLM) × 2 (IV2 - Disclosed Label: Human professionals vs. Artificial Intelligence) × 3 (IV3 - Information Type: General vs. Symptom vs. Treatment) factorial design to explore people's perceived trust in online health information. The source of health information (IV1) was set as a between-subjects variable to explore whether people have different trust perceptions based on the source (human professionals vs. LLM), which might inherently present information in distinct styles. A within-subject design for the source could introduce biases in perceived quality and trustworthiness due to these stylistic differences. Additionally, using a between-subjects design for the source helps isolate the effect of labeling (IV2), making the findings clearer and more robust. Conversely, for the label of the source (IV2) and the type of health information (IV3), we opted for a within-subjects design to allow direct comparisons of trust perception across different labels and types while keeping the source uniform for each participant. This approach reduced individual variability, ensuring a clearer separation of source effects on trust variances while enabling robust analysis of influences from labeling and types of health information. Therefore, during the completion of the survey, each participant read the information either generated by human professionals or LLMs, and each of them experienced six distinct conditions.

Health information.

Sets of health information (question and answer pairs) from human professionals were selected from an open-sourced dataset [395] due to its comprehensive and diverse range of health-related questions, authored by certified professionals. It ensured the reliability and authenticity of the information used for comparison in this work. To produce comparable LLM-generated responses, we used the Generative Pre-trained Transformer 4 (GPT-4) model [196] (version: "gpt-4-0125-preview" through official API) and prompted it with the selected health question and accompanying instruction (e.g., " Health question: [question]. Please give the answer to the above question within [wordcount] words?") to generate answers of similar length to those from human professionals. To ensure content accuracy and mitigate potential misinformation, all LLM-generated responses were reviewed by the researchers for correctness and alignment with existing health guidelines prior to inclusion in the study. Any responses identified as containing errors or inaccuracies were corrected to maintain consistency and fairness in comparisons between sources.

The health information falls into three categories: **General information**: provides answers to general health topics (e.g. "Do you have information about weight control?"); **Symptoms-related information**: focuses on symptoms and potential diagnoses (e.g. "What are the symptoms of burns?"); **Treatment-related information**: provides treatment options for specific conditions (e.g. "What to do for burns?").

Twenty-five questions were selected from each category resulting in a question set with 75 questions in total, ensuring a comprehensive representation of individual health questions.

Measures

Demographics and prior experience. In the pre-survey, we collected participants' demographic information (age, gender, education, occupation) and their experience in online health information seeking, using two questions: "How often do you search for health information online?" rated on a 5-point Likert scale from Never to Daily; and "How long have you been using online sources for health information searching?" with options ranging from Less than 1 year to More than 10 years.

Propensity of trust in technology (PPT) [300] was used to assess inherent trust in technology before participants read the health information. It consists of 6 items looking at people's general trust in technology (e.g. "I think it's a good idea to rely on technology for help"). All items were scored on a 5-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) (Cronbach's $\alpha = 0.71$).

eHealth and AI literacy As part of the pre-survey, we also measured participants' literacy on eHealth and AI separately using two adapted questionnaires from eHEALS: The eHealth Literacy Scale [396] and MAILS - Meta AI literacy scale [307]. All the items were scored from 1 (Strongly Disagree) to 5 (Strongly Agree). The adapted measure for eHealth literacy has eight items with an example as "I know where to find helpful health resources on the Internet" (Cronbach's $\alpha = 0.88$), and the adapted measure for AI literacy has ten items with an example item as "I can distinguish if I interact with an AI or a real human" (Cronbach's $\alpha = 0.76$).

Trust of online health information [254, 301] (**Trust Score**) During the formal study, participants completed the trust of online health information questionnaire to rate their trust level after reading each set of health information. It consists of 13 items (e.g. "The information appears to be objective."), each rated on a 5-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) (Cronbach's $\alpha = 0.92$). We aggregated and calculated the average value of all 13 items to obtain our perceived **trust score**. We use this score for further analysis throughout our work.

Post-survey: three open-ended questions At the end of the survey, participants were asked to reflect on their trust perceptions through three open-ended questions. These questions explored their views on (a) general trust in LLM-generated information versus information from human professionals, (b) how they assess the credibility of online information, and (c) how the labeling of the health information source influences their perceived trust.

Participants

Participants were recruited through the online crowd-sourcing platforms Prolific [397] and institute recruitment systems. Our inclusion criteria included individuals over the age of 18 who are fluent in English, and they must have passed the attention check. A power analysis conducted with G*Power 3.1 [207] for a mixed-factor ANOVA design indicated that a minimum of 76 participants would be required to detect a small effect size (f=0.15), with an alpha level of 0.05 and a power of 0.95.

A summary of participants' demographic information is shown in Table 6.1. 142 participants (N=142) were recruited (F=83, M=58, NB=1), with 90.9% falling in the 18-34 age bracket. Regarding educational backgrounds, 47.2% had undergraduate degrees, and 35.9% held postgraduate qualifications. As for online health information-seeking

Demographic	Categories	Numbers of Participants (%)
Gender		(N=142)
	Female	83 (58.5%)
	Male	58 (40.8%)
	Non-binary	1 (0.7%)
Age		
	18-24	91 (64.1%)
	25-34	38 (26.8%)
	35-44	9 (6.3%)
	45-54	2 (1.4%)
	65+	2 (1.4%)
Education		
	High school degree or equivalent	24 (16.9%)
	Bachelor ’ s degree	67 (47.2%)
	Master ’ s degree	49 (34.5%)
	Doctorate or higher	2 (1.4%)
Professional Domain		
	Health and Medical Science	17 (12.0%)
	Science, Technology, Engineering, and Mathematics (STEM)	35 (24.6%)
	Business, Economics, and Law	35 (24.6%)
	Arts, Culture and Entertainment	19 (13.4%)
	Government and Public Sector	3 (2.1%)
	Education	3 (2.1%)
	Other	30 (21.1%)
Frequency of online health information seeking		
	Rarely	27 (19.0%)
	Sometimes	77 (54.2%)
	Often	31 (21.8%)
	Always	7 (4.9%)
Duration of online health information seeking		
	Less than 1 year	4 (2.8%)
	1-3 years	24 (16.9%)
	3-5 years	51 (35.9%)
	5-10 years	45 (31.7%)
	More than 10 years	18 (12.7%)

Table 6.1: Characteristics of participants in the online survey.

experience, 26.7% frequently used online sources, 54.2% occasionally searched online, and 19.0% rarely used online resources.

Procedure

The study design and procedure are outlined in Fig 6.1(a). Participants were first provided with detailed information about the study and gave informed consent in line with institutional guidelines. They provided demographic information and their experiences with online health information seeking. A total of 75 health questions were used in the online survey, divided evenly into three categories: general health, symptom-related, and treatment-related (25 each) (Sec 6.3.1 “Health information”). For each participant, six Q&A pairs were shown: two randomly selected from each category. The survey study used a between-subjects design for the source of the information (AI- vs. human-generated) and a within-subjects design for the label (AI- vs. human-labeled). Both source and label orderings were counterbalanced across participants to mitigate order effects. After reviewing each Q&A pair, participants rated their perceived trust in the information.

Afterward, participants completed a post-survey comprising three open-ended questions about their perceptions of the information source and its labeling. Participation was voluntary and participants were monetarily compensated for a 30-minute session. To ensure we avoided bots in our responses, we included an additional attention check where respondents needed to select a specific response to one question. Our study received approval from our institute's ethics and data protection committee.

Data analysis

We conducted quantitative analyses to examine how the types of health questions, information sources, and labeling of sources influence trust perception in online health information. Initially, we confirmed the data's suitability for parametric tests by performing the Shapiro-Wilk test [208] for normality and Bartlett's test [209] for homogeneity of variance; neither assumption was violated. Next, we performed a mixed model, i.e., three-way mixed ANOVA [398] to investigate differences in trust perceptions based on information sources, disclosed labels, and types of information. Since only one ANOVA was conducted, no correction for multiple tests was applied. Following, post-hoc pairwise comparisons were conducted using t-tests with False Discovery Rate (FDR) correction [399] to examine differences in trust between each pair of label and source combinations. To explore the relationship across variables, we also conducted the Pearson correlation analyses [311] on two subsets of the data: one with human-sourced sources and the other with LLM-sourced sources (between-subject independent variable). Bonferroni correction [212] was applied to account for multiple comparisons in both correlation analyses.

We conducted an inductive content analysis [214] on the responses to three open-ended questions, focusing on identifying underlying themes that explain trust rather than counting frequencies. In the first stage, the first two authors created an initial set of codes using the qualitative analysis software ATLAS.ti [215]. This initial codebook looked at respondents' varying perceived trust in AI and human professionals, their reasons for trusting or distrusting, and how they typically evaluate the credibility and trustworthiness of information. Following this, both coders independently open-coded the responses, remaining open to new observations and emerging codes. Similar codes were merged, unclear ones were refined, and earlier responses were recoded as needed. As the analysis progressed, recurring factors emerged across different questions, allowing us to develop common themes that spanned all three sets of responses.

6.3.2 Quantitative Findings

Descriptive statistics

	Measures	**Mean**	**SD**
Pre-survey	Propensity of trust in AI technology (PPT)	3.85 / 5	.72
	eHealth literacy	3.62 / 5	.87
	AI literacy	3.81 / 5	.92
	Conditions	**Mean**	**SD**
Trust score	Source (Human) & Label (Human)	4.01 / 5	.45
	Source (Human) & Label (AI)	3.76 / 5	.49
	Source (LLM) & Label (Human)	4.07 / 5	.47
	Source (LLM) & Label (AI)	3.87 / 5	.44
	Source (Human), regardless of Label	3.89 / 5	.84
	Source (LLM), regardless of Label	3.97 / 5	.81
	Label (Human), regardless of Source	4.04 / 5	.46
	Label (AI), regardless of Source	3.82 / 5	.47

Table 6.2: Descriptive statistics of the online survey.

As shown in Table 6.2, participants demonstrated a positive propensity to trust in technology, with an average score of 3.85 (SD=.72), indicating a positive attitude towards technology. The average eHealth literacy score was 3.62 (SD=.87), indicating that participants are relatively capable of using online health resources. AI literacy was also high, with an average score of 3.81 (SD=.92), reflecting a favorable understanding of AI technology.

In terms of trust perception, the trust scores (based on the aggregate Trust Score described in Sec 6.3.1) varied depending on the source and label of the information. For information both sourced from and labeled as human, the average trust score was 4.01 (SD=45). When the information was sourced from humans but labeled as AI, the trust score decreased significantly to 3.76 (SD=.64). In contrast, information sourced from LLM but labeled as human received the highest trust score of 4.07 (SD=.47), while information sourced from AI and labeled as LLM had a trust score of 3.87 (SD =.44). These findings highlighted the ways in which both the source and labeling of information can impact trust perceptions, with a clear indication that labeling of the sources plays a role in shaping trust, potentially even more than the actual source of the information.

Our mixed model analysis compared differences in trust levels among the source, label, and health information types. Findings are shown in Table 6.3 and Fig 6.2, and together highlight how people perceive and trust health information manipulated by sources and labels.

Outcomes	Conditions	F Statistic	P-value	Effect size (η^2)
Trust score	Source (Human vs. LLM)	2.27	.024 **	.14 (medium)
	Label (Human vs. AI)	-6.50	< .001 **	-.39 (medium)
	Type of health information	0.67	.505	.03 (small)

Table 6.3: Results from the three-way mixed ANOVA analysis on the trust score without data correction. (**p<.01, *p<.05)

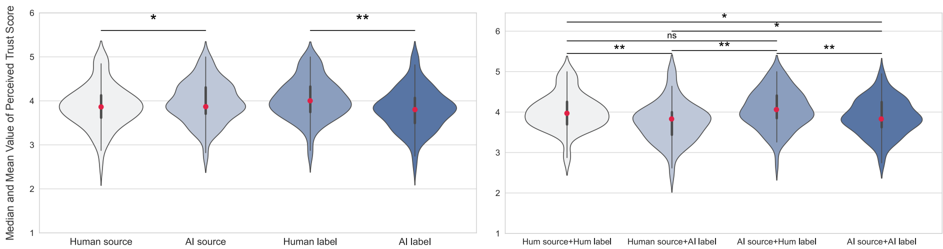


Figure 6.2: Left: Perceived trust score in information by sources regardless of labels, and by labels regardless of source from the three-way mixed ANOVA without correction. **Right:** Post hoc pairwise comparisons on perceived trust score based on different source and label conditions using t-test with False Discovery Rate (FDR) correction. Each plot shows the score density (width), with the red dot indicating the mean, the black line as the median, and thick bars representing the interquartile range (IQR). Horizontal lines indicate significance (**p<.01, *p<.05, "ns": no significance).

Participants gave higher trust levels to health information sourced from LLM than from human professionals.

The impact of the information source (human professionals vs. LLM) on trust in health information was analyzed by a three-way mixed ANOVA. The results showed significant differences in trust levels between sources: statistics=2.27, p=.024, effect size=.14. This suggests that information sources significantly influence overall trust in health information. Specifically, participants reported trusting information from LLM more than human professionals, with an average trust score for LLM-sourced information of 3.97 (SD=.81), compared to 3.89 (SD=.84) for information from human professionals. Although the difference was modest, it indicated increasing user acceptance of LLMs as credible sources of health information. However, perceived trust does not imply factual accuracy. These findings reflected shifting attitudes toward AI-generated content but do not confirm its factual reliability compared to health professional advice.

Participants gave higher trust ratings to health information labeled as from human professionals compared to labeled as from AI.

Except for the factor of 'source', the labeling of information sources influenced trust perception significantly. Participants perceived significantly lower trust in health information labeled as from AI compared to that labeled as from human professionals, as indicated by a mixed model ANOVA (statistics=-6.50, $p<.001$, effect size=-.39), with an average trust score for information labeled as from human professionals of 4.04 (SD=.46) and 3.82 (SD=.47) for information labeled as from AI. We also observed no significant difference in trust between human-labeled information from human sources (M=4.01, SD=.45) and LLM sources (M=4.07, SD=.47). These results suggested that while LLM-generated information is generally trusted, the perceived trust still leans in favor of human-associated information when directly compared.

Type of health information does not affect participants' trust level in information.

Additionally, we explored how trust varied across different categories of health information. There was no significant effect found (statistics=0.67, $p=.505$, effect size=.05). This suggests that the type of health question does not influence people's trust levels in health information. The interaction effect between the label of the information source and category of information was not significant as well (statistics=-.51, $p=.613$, effect size=-.15) This implied that the influence of labeling on trust does not vary across different types of health information.

Correlation analysis

Given that the mixed ANOVA indicated no significant effect of the type of health information on the trust perceptions, the repeated measures were averaged into a single observation for each participant. This simplification allowed us to conduct a Pearson correlation analysis [311] to examine the general relationships between key variables in the online survey. The results, illustrated in Fig 6.3, revealed distinct patterns of trust in health information from different sources. For information sourced from human professionals, trust in human-labeled information showed a moderate positive correlation with trust in AI-labeled information ($r(142)=0.47$, $p<0.01$). However, other relationships, such as those involving eHealth literacy and AI literacy, exhibited weak or negligible correlations. In contrast, for information sourced from LLMs, we observed stronger correlations across multiple variables. Trust in human-labeled information showed a strong positive correlation with trust in AI-labeled information ($r(142)=0.65, p<0.01$), AI literacy ($r(142)=0.41, p<0.01$), and the propensity of trust in AI ($r(142)=0.37, p<0.05$). Additionally, the propensity of trust in AI correlated to trust in AI-labeled information ($r(142)=0.50, p<0.01$), eHealth literacy ($r(142)=0.42, p<0.01$), and AI literacy ($r(142)=0.30, p<0.01$) AI literacy positively correlated with eHealth literacy ($r(142)=0.33, p<0.01$). These results highlighted a consistent influence of labeling on participants' trust across different sources.

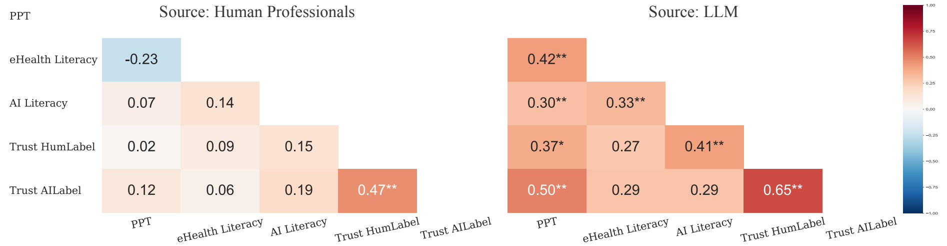


Figure 6.3: Pearson correlation with Bonferroni correction among the key variables in the online survey.(**p<.01, *p<.05) Note: "HumLabel": information with human label regardless of the actual source. "AILabel": information with AI label regardless of the actual source.

6.3.3 Qualitative Findings

We received a total of 426 free-text responses (142 for each question). In this section, we present our findings with four themes. We found that participants' trust in AI versus humans is shaped by their inherent trust predispositions and their perceived source of knowledge. Additionally, participants valued human consciousness as a factor contributing to greater trust, and the presentation of information also influenced their trust.

Predisposition toward AI and humans influences trust

Survey respondents demonstrated a predisposition to trust either AI or humans, independent of the content or source of the information. However, there were individual differences in this inclination. Some respondents were optimistic about AI technology, regularly using and trusting AI in their daily lives. They perceived no difference in reliability between AI and human professionals, and some even trusted AI more. Conversely, some respondents expressed significant reservations about AI, doubting its readiness to address serious topics, especially in sensitive fields like healthcare. One respondent noted, *"I don't trust AI, and the quick push in its advancements is dangerous; at the very least, it should be limited in specific fields such as health."* Privacy concerns and the risks of AI-driven health advice reinforced such skepticism, leading to more critical evaluation of AI recommendations. This underlying predisposition toward AI or human professionals also shaped respondents' views on labeling. Some participants expressed a preference for human-labeled content, with one stating, *"AI label makes me trust it less and view the information more critically than if it came from a human professional."* However, not all respondents allowed their predispositions to dictate their trust. Others placed less emphasis on labels, focusing instead on verifying information from multiple perspectives rather than relying solely on the source. As one respondent explained, *"The label doesn't affect* *how I interact with it, and my trust wouldn't be based solely on the label."*

Perceived source of knowledge influences trust

Survey respondents' trust in AI or human professionals was shaped by their perceptions of where each derives its knowledge. One respondent explained, *"I would trust a human professional more, since he has learned factual information in school. An AI has learned from multiple sources online, not only factual ones, so that is why I would trust them a bit less."*. In contrast, some respondents believed that AI can learn from *"more databases and the most important points that all research brought up"*, potentially making it more knowledgeable than a single human expert. These differing views on the origins of human and AI's knowledge contributed to varying levels of trust. Some respondents took a more balanced stance, recognizing that both AI and human professionals are susceptible to biases and errors. As one respondent commented, *"While information from a human professional may need correction due to incomplete knowledge, information from AI might contain errors due to gaps in its training data."* Consequently, many respondents shared that they would evaluate both sources of information with equal care, relying on their own experiences to evaluate the content's credibility. Additionally, some respondents expressed a preference for combining information sources, such as cross-checking information or using AI as a complementary tool to support human decision-making.

6

The human touch builds greater trust than AI

Survey respondents highlighted that, due to the absence of consciousness and empathy in AI, they trusted human professionals more, particularly in healthcare contexts. Many respondents emphasized that AI lacks the ability to evaluate information with awareness. As one respondent commented, *"Unlike human, AI doesn't know the difference between good or bad quality."* In contrast, many respondents emphasized that human professionals have *"years of medical education and experience with real-life cases"* to inform their decisions, something that AI cannot replace despite its access to vast information. This absence of consciousness made respondents very skeptical about AI's capability to offer reliable health advice. The issue extended beyond decision-making to interpersonal interactions. Respondents valued the sense of responsibility and ethical obligation that human professionals carry, with one noting, *"I trust the information from the human professional more because they are human and have moral and professional obligations about not giving misinformation."* Additionally, human-to-human interaction offered a sense of personalized care, making respondents feel their symptoms are better understood. In contrast, AI lacked this human touch, and its absence of empathy and accountability led respondents to trust it less.

Presentation of information influences trust

Information presentation was highlighted as an advantage of AI, which increased respondents' trust. They mentioned that when evaluating health information, factors such as the design of the user interface, the length of the information, the visible publication date, and the clarity of language were important. Compared to human professionals, AI was often perceived as providing simpler, more structured, and user-friendly information. Respondents appreciated that AI's answers were clearly explained and easy to understand. Additionally, the objective tone of AI responses further boosted respondents' trust. These elements collectively enhanced AI's explainability. As one respondent noted, *"When I receive information from a human professional, I expect it to contain more academic language, which is harder to understand and less explanatory. Information from AI, however, uses simpler words and is easier to understand."*

6.4 Study 2: Laboratory Study

Study 1 demonstrated that the factors of actual source and disclosed label both affect people's perceived trust (self-reported) in health information. To further understand the process and user behaviors involved in forming trust perceptions, we conducted an in-person experiment. This study explored how health information from different sources and labels affects people's behavioral and physiological states.

6.4.1 Study Methods

Design

Similarly to the online survey study, we utilized a within-subjects 2 (IV1 - Information Source: Human Professional vs. LLM) x 2 (IV2 - Disclosed Label: Human Professional vs. Artificial Intelligence) x 3 (IV3 - Information Type: General vs. Symptom vs. Treatment) factorial design tested in a controlled, laboratory environment (as shown in Fig 6.4). Different from Study 1, participants experienced all 12 distinct conditions for this in-person lab experiment, enabling direct comparisons between human and LLM-generated health information. We opted for a within-subjects design for all independent variables to facilitate a nuanced analysis of participants' behavioral and physiological responses across different conditions. Specifically, for the source of information (IV1), we aimed to observe whether participants exhibited different behavioral (e.g., gaze patterns) and physiological (e.g., heart rate, skin conductance) signals when reading information attributed to human versus LLM sources. While these sources may differ in presentation styles, it is also possible that participants' trust per-

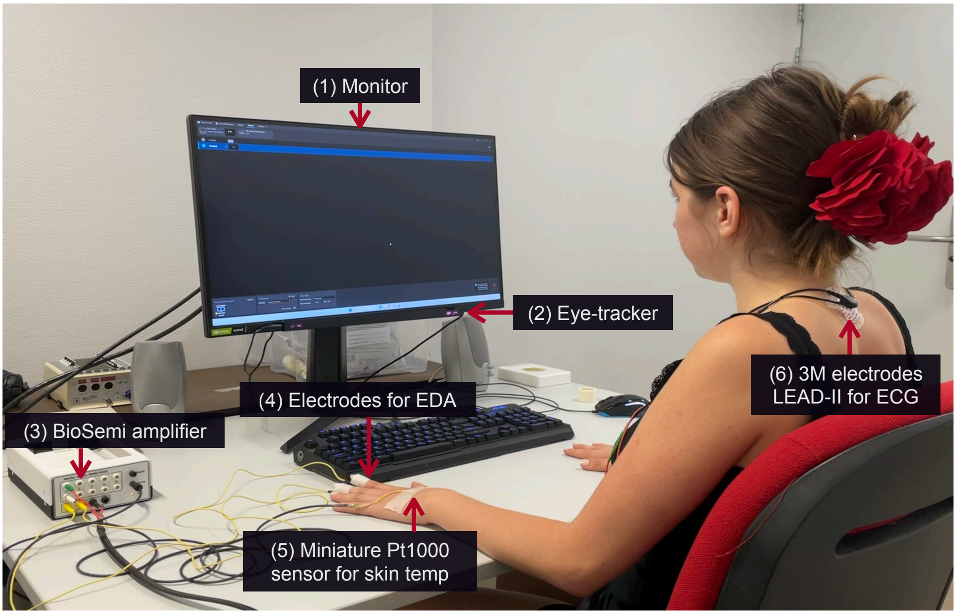


Figure 6.4: The hardware setup for presenting the text stimulus and collecting physiological signals, eye movement, and pupil dilation.

ceptions were influenced more by their belief about the source of the text (human vs. AI) rather than the actual content or style. A within-subject design was critical for disentangling these effects, as it allowed each participant to serve as their own control, reducing variability across conditions and enabling a clearer examination of these factors. Participants rated their perceived level of trust for each set of health information while their eye-tracking data (gaze positions and pupil diameter) and physiological responses (ECG: Beats Per Minute (BPM), Beat-to-Beat Interval (BPI), Root Mean Square of Successive Differences (RMSSD); EDA: Skin Conductance Level (SCL) and Response (SCR); Skin Temperature) were recorded throughout the tasks.

To address our second research question, we explored whether behavioral and physiological signals can be used to understand and predict trust perceptions toward human and AI-generated personal health information. Therefore, we set up two prediction tasks that make use of our sensed data: (1) predicting participants' perceived trust in health information by both regression on perceived trust scores and binary classification on trust level (high vs. low); and (2) classifying the actual source of the health information.

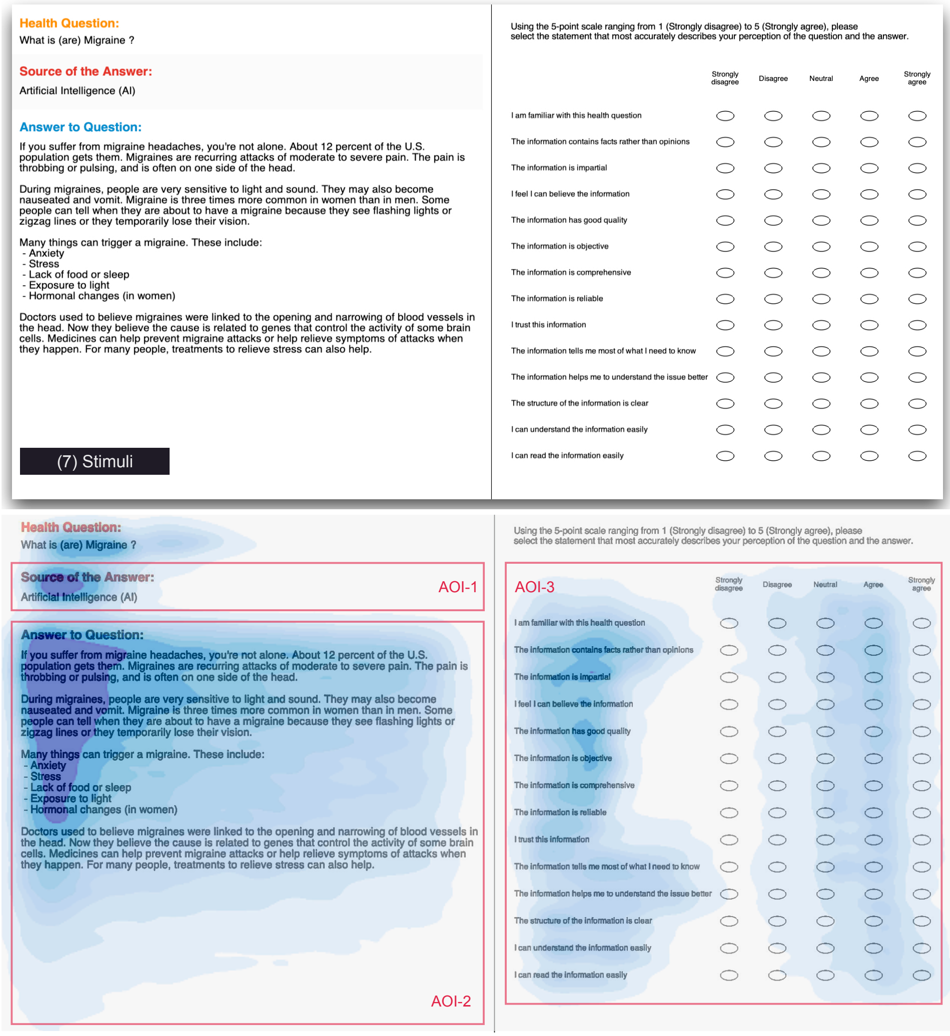


Figure 6.5: Top: An example of text stimulus displayed on the monitor. **Bottom:** Heatmap of the gaze points on stimuli. Three AOIs are predefined: AOI-1 is the area for presenting disclosed label; AOI-2 is the area for presenting health information; AOI-3 is the area to rate the perceived trust in health information.

Stimuli and apparatus

We developed a web interface that displays the health information (question and answer pair) and the questionnaires for participants to rate their trust scores (see Fig 6.5). The health information was identical to the material used in Study 1, as described in Section 6.3.1. Each set of health information was labeled as being generated either by "Human Professionals" or "Artificial Intelligence", regardless of the actual source.

We used a PHILIPS (full HD, 1920*1080, 100 Hz) monitor to display the stimuli. The eye movements and pupil diameter (PD) data were recorded by Tobii Pro Fusion eye tracker. The remote eye tracker was attached to the bottom of the monitor and connected to a computer (Windows, Intel core i5, 16GB RAM) running the Tobii Pro Lab software [400].

Physiological signals, including ECG, EDA, and skin temperature, were measured using a BioSemi amplifier [401] (as shown in Fig 6.4). ECG was captured through a disposable 3M Red Dot in LEAD-II configuration, EDA was measured with electrodes attached to fingers, and skin temperature was monitored with a miniature Pt1000 sensor, all at a 24-bit resolution and 1000 S/s sampling rate. These data were collected using software FysioRecorder version 2.1 [401].

Data recording was initiated through a central recording application developed in PsychoPy [402], connecting to sensors via IP addresses to simultaneously capture synchronized ECG, EDA, skin temperature, and eye-tracking signals.

Self-reported measures

We collected several self-reported measures, consistent with those used in Study 1 described in Sec 6.3.1. These included demographics, prior experience with online health information and AI, the propensity to trust technology (PPT), eHealth, and AI literacy. Additionally, we assessed the perceived reliability of AI and human professionals using a single item for each: "How reliable do you find AI/Human Professionals?" Responses were captured on a 5-point Likert scale, ranging from 1 (Not at all) to 5 (Extremely). They were collected before the formal reading task.

During the reading task, we repeatedly measured the participants' 1) familiarity level with each given health question and 2) their perceived trust score in health information [254, 301], after they completed each stimulus.

Machine learning: setup and approach

We performed binary classification to predict information sources and applied both regression and classification (i.e., binary and three-class classification) for trust scores. The perceived trust score (see Sec 6.4.1), as an aggregate numerical rating based on the Trust of online health information questionnaire [254, 301], naturally lends itself to regression. However, this approach can be challenging to interpret given that trust is an aggregate and overall complex construct. On the other hand, trust classification simplifies interpretation but introduces an arbitrary split between high and low trust levels. To address this, we pre-processed the original trust scores into high and low categories using the median value as a threshold for binary classification. For the three-class classification, we divided the trust scores into low, medium, and high categories based on tertiles, creating balanced splits that account for the distribution of scores.

We used several common machine learning algorithms as suggested in prior research [393], including single models (i.e., Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Linear Regression, Ridge Regression, Random Forest-based Regression), and ensemble methods (i.e., boosting, voting, stacking and bagging). Models were built using hand-crafted gaze features (i.e., fixations, saccades, pupil diameter) and physiological signals (i.e., BPM, BPI, RMSSD, SCL, SCR, and skin temperature).

We experimented with three feature sets: Gaze-only, Physiology-only, and Gaze + Physiology. These sets trained and evaluated the selected models to determine how effectively they could predict participants' perceived trust scores and classify the source of information. We only considered user-independent models to ensure that any predictions generalize across all participants, despite well-known challenges in generalizing using peripheral physiological features [403]. Additionally, we set the "random state" [404] parameter to ensure result consistency and used the "grid search" [405] technique to find the optimal hyper-parameters of the models. The performance of the regression models (for trust score prediction) was evaluated by Mean Squared Error (MSE) and Coefficient of Determination (R^2). The performance of the classification models (for trust level and information sources prediction) was assessed through 10-fold cross-validation [406], with the Macro-F1 [407] score as the average of these validations.

Participants

For the in-person experiment, we used the same inclusion criteria as in Study 1 (age above 18 who are fluent in English language understanding). Participants were recruited through the institute's recruitment systems. A power analysis using G*Power 3.1 [207] for a within factors ANOVA indicated that at least 28 participants were required to detect a medium effect size seen in Study 1 (0.24), with an alpha level of 0.05 and a power of 80%.

The characteristic information of the participants is listed in Table 6.4. Forty participants (N = 40) were enrolled (F = 23, M = 16, NB = 1) aged between 18 to 65+ years, with 92.5% falling in the 18-34 age range. Regarding online health information-seeking experience, 22.5% frequently or always used online sources, 62.5% occasionally searched online, and 15.0% rarely used online resources. For the frequency of AI usage, 60.0% frequently or always used AI tools, 22.5% occasionally used AI, and 17.5% rarely or never used AI.

Demographic	**Categories**	**Numbers of Participants (%)**
Gender		(N=40)
	Female	23 (57.5%)
	Male	16 (40.0%)
	Non-binary	1 (2.5%)
Age		
	18-24	23 (57.5%)
	25-34	14 (35.0%)
	35-44	1 (2.5%)
	45-54	1 (2.5%)
	65+	1 (2.5%)
Education		
	Bachelor' s degree	18 (45.0%)
	Master' s degree	17 (42.5%)
	Doctorate or higher	5 (12.5%)
Professional Domain		
	Health and Medical Science	2 (5.0%)
	Science, Technology, Engineering, and Mathematics (STEM)	11 (27.5%)
	Business, Economics, and Law	8 (20.0%)
	Communication, Arts, Culture, and Entertainment	7 (17.5%)
	Education and Social Science	7 (17.5%)
	Other	5 (12.5%)
Frequency of online health information seeking		
	Rarely	6 (15.0%)
	Sometimes	25 (62.5%)
	Often	7 (17.5%)
	Always	2 (5.0%)
Frequency of using AI tools		
	Never	2 (5.0%)
	Rarely	5 (12.5%)
	Sometimes	9 (22.5%)
	Often	18 (45.0%)
	Always	6 (15.0%)
Duration of online health information seeking		
	Less than 1 year	4 (2.8%)
	1-3 years	24 (17.0%)
	3-5 years	50 (35.5%)
	5-10 years	45 (31.9%)
	More than 10 years	18 (12.8%)

Table 6.4: Characteristics of participants in the lab study.

Study procedure

Each participant was invited to the institute for a single session at the lab. The researcher first provided an overview of the study and task details, after which participants gave informed consent before the lab session. During the pre-survey, participants provided their demographic information (age, gender, occupation) and their experiences with online health information search and interactions with AI.

Upon completing the pre-survey and successfully calibrating the sensors, participants began the formal reading task. During the reading task, each participant reviewed 12 sets of health information: six were labeled as from human professionals and six as from AI, regardless of the actual source. Sources and labels were counterbalanced to minimize order effects. The entire session lasted approximately 60 minutes, and participants were rewarded with €10 for participating. The study received approval from our institute's ethics and data protection committee. The procedure of the lab study is detailed in Fig 6.1(b).

Data pre-processing

Self-reported Trust Scores. To assess how factors such as information source, labeling, and information type affect trust in online health information, we first checked the suitability of the data for statistical analysis. A Shapiro-Wilk test [208] confirmed that the self-reported trust scores deviated from a normal distribution. Therefore, we applied generalized estimating equations (GEE) [210] to analyze trust differences across information sources and labels, because of its robustness to violations of normality and flexibility to handle repeated ordinal measures. Additionally, we conducted Spearman correlation analyses [213] with Bonferroni corrections to explore relationships among the variables. Consistent with Study 1, and given that the GEE results (Table 6.6) indicated no significant interaction effects between the independent variables of source and labeling, we averaged the repeated measures for each participant into a single observation across conditions. This simplification allowed us to focus on the key exploratory relationships while maintaining analytical clarity.

Eye Tracking Data Processing. Raw eye-tracking data were extracted from the eye tracker (Tobii Pro Fusion) using Tobii Pro Lab software [400]), and time-synchronized with stimuli. As shown in Fig 6.5, there are three Areas of Interest (AOIs): AOI-1 (disclosed label of source), AOI-2 (health information), and AOI-3 (rating scale). We chose fixation threshold of 30° for velocity and 60 ms for duration, as suggested by online information reading task [408]. Gaze features including gaze duration, fixation (count and duration), saccade (count and length), and pupil diameter were calculated to understand how users read the information. We excluded participants whose gaze accuracy fell below 90%, resulting in 38 participants' eye-tracking data being retained for further analysis. After transforming data through Aligned Ranked Transformation (ART) [409], we confirmed the non-normality of eye tracking data with the Shapiro-Wilk test.

Physiological Signal Processing. Physiological signal was processed using Vsrrp98 software (v13.1.4) [401], following the practise in prior research [61, 62]. Key physiological features derived from the raw ECG data using interval-to-interval window size included BPM, BPI, and the main HRV metrics - the root mean square of successive differences (RMSSD). For EDA data, we used the continuous decomposition analysis method [410] to separate it into the tonic SCL and phasic SCR components, then calculated the mean SCL and SCR values, SCR count. Skin temperature readings were screened for any abnormal responses. We excluded SCL and SCR data when more than 4 out of 12 stimuli have values lower than .01μS or exceeded 50μS, as these readings likely resulted from loss recording or movement artifacts. As a result, we retained data from 34 participants for SCL and SCR analysis, and 40 participants for ECG and skin temperature analysis. Following preprocessing, we used the Shapiro-Wilk test to assess normality, revealing that all physiological features were not normally distributed.

Given the exploratory nature of our investigation and the presence of multiple comparisons, we applied appropriate corrections based on the type of data. First, self-reported data were analyzed using a single GEE test, thus no multiple comparison correction was necessary. Second, for eye-tracking data, where multiple tests were conducted for different features, we applied False Discovery Rate (FDR) correction [399] to control for potential inflation of Type I errors. Third, for physiological data, no multiple comparison correction was applied because most of the physiological features (e.g., RMSSD, ECG, EDA) were uncorrelated, as confirmed by correlation analysis, and each feature was analyzed independently. This approach reflected our goal of treating these features as distinct, non-overlapping measures, without assuming that they influence each other.

6

6.4.2 Findings

Descriptive statistics

	Measures	**Mean**	**SD**
Pre-survey	Familiarity of AI	3.58	.96
	Perceived Reliability of AI	3.13	.61
	Perceived Reliability of Human Professionals	3.78	.53
	Propensity of Trust (PPT)	3.54	.33
	eHealth literacy	3.69	.25
	AI literacy	3.78	.20
	Conditions	**Mean**	**SD**
Trust score	Source (Human) & Label (Human)	3.67	.63
	Source (Human) & Label (AI)	3.56	.64
	Source (LLM) & Label (Human)	3.92	.56
	Source (LLM) & Label (AI)	3.78	.63
	Source (Human), regardless of Label	3.62	.64
	Source (LLM), regardless of Label	3.85	.60
	Label (Human), regardless of Source	3.80	.61
	Label (AI), regardless of Source	3.67	.65

Table 6.5: Descriptive statistics of the lab study.

As shown in Table 6.5, participants demonstrated a generally positive attitude toward technology, with an average trust in technology score of 3.36 (SD=.23). Their eHealth literacy averaged 3.69 (SD=.25), indicating proficiency in searching for digital health information. AI literacy was even higher, with an average score of 3.78 (SD=.20), suggesting a strong understanding of AI and its applications.

Regarding perceived trust, the lab study results closely mirrored those of the online survey, despite being based on separate participant samples and independently collected data. The self-reported trust scores from the lab study varied depending on both the source and the labeling of the health information. Information both sourced from and labeled as from human professionals had an average trust score of 3.67 (SD=.63). When human-sourced information was labeled as AI, the score slightly decreased to 3.56 (SD=.64). LLM-sourced information labeled as from human received the highest trust score of 3.92 (SD=.56), while information sourced from LLM and labeled as from AI had a trust score of 3.78 (SD=.63). Overall, participants reported higher trust in LLM-sourced information (M=3.85, SD=.60) than in human-sourced information (M=3.62, SD=.64), echoing the trend observed in the online survey and indicating a growing acceptance of AI (i.e., LLM) in health contexts. However, information labeled as coming from human professionals was trusted more (M=3.80, SD=.61) than that labeled as AI (M=3.67, SD=.65), suggesting that labeling plays an influential role in trust formation, potentially even more than the actual source. These findings reinforced the patterns found in the online survey while providing additional validity through the lab sessions.

Analysis of self-reported trust

Outcomes	Conditions	Coefficient (β)	P-value	Effect ($Std.$)
Trust score	Source (Human vs. LLM)	.22	< .001 **	.35 (medium)
	Label (Human vs. AI)	-.15	.01 **	.23 (medium)
	Source $*$ Label	.03	.71	.05 (small)

Table 6.6: Results from the GEE analysis on the self-reported trust score. (**p<.01, *p<.05)

Table 6.6 presents the results from the GEE analysis on self-reported trust scores from the lab study. Consistent with the online survey, both the source and the label significantly impacted trust perceptions. Fig 6.6 further illustrates the same pattern, echoing the online survey results. Trust was highest for LLM-sourced information labeled as human and lowest for human-sourced information labeled as AI.

The analyses first revealed a significant effect of information source on trust, with a coefficient of 0.22 ($p < 0.01$), indicating that LLM-sourced information was generally trusted more than human-sourced information, i.e., without knowledge of the actual source. This suggested that the source of information is crucial in shaping

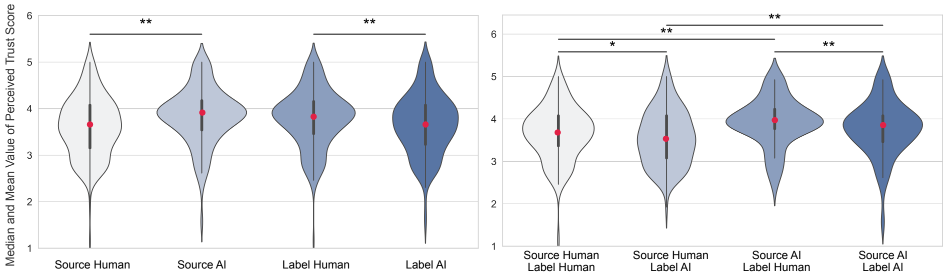


Figure 6.6: Left: Perceived trust score in information by sources regardless of labels, and by labels regardless of sources. **Right:** Perceived trust score based on different source and label conditions. Each plot shows score density (width), with the red dot as the mean, the black line as the median, and thick bars denoting the interquartile range (IQR). Horizontal lines indicate significant pairwise differences (**p<.01, *p<.05).

trust, as AI-generated content may be perceived as more structured and objective than human-authored content. While the raw coefficient represented a modest change of 0.22 points on the 5-point Likert scale, the corresponding standardized effect size ($Std.\beta$=0.35) was classified as medium. This reflected the bounded nature of the Likert scale, where even small raw differences can indicate meaningful relationships due to the relatively low variability in responses. Thus, the medium effect size underscored the practical relevance of the findings despite the small-scale differences.

Labeling also significantly impacted trust, with a coefficient of -0.15 ($p = 0.01$), meaning information labeled as AI was trusted less than when labeled as human professionals. The negative coefficient suggested a preference for human-labeled information, as participants may associate human endorsement with greater credibility. Similarly, while the raw change (-0.15) was modest, the standardized effect size ($Std.\beta$=0.23) reflected a medium effect, emphasizing that the impact of labeling, though subtle on the scale, has measurable and meaningful implications for trust perception.

Notably, the interaction between source and label was not significant with a coefficient of 0.03 (p = 0.71), indicating that the combined influence of source and label does not affect trust beyond their individual effects. The small standardized effect size ($Std.\beta$=0.05) confirmed that this interaction effect is negligible.

Analysis of eye movement data

The results of GEE analysis [210] on eye tracking data are detailed in Table 6.7, showing varied eye movement patterns. In AOI-1 (label area), fixation (count and duration), saccade count, and pupil diameter of fixation showed significant differences by information sources and labels. In AOI-2 (main information area), fixation duration

Outcomes	AOI	Conditions	Coefficient (β)	P-value (orig)	P-value (corr)	Effect ($Std.\beta$)
Fixation Count	AOI-1	Source	-4.02	.033	**.079**	.22 (medium)
		Label	-4.54	.015	**.055**	.25 (medium)
		Source × Label	6.01	.082	.150	.33 (medium)
	AOI-2	Source	0.61	.962	.962	.00 (small)
		Label	-3.23	.827	.910	.02 (small)
		Source × Label	34.61	.036	**.079**	.26 (medium)
	AOI-3	Source	-8.14	.198	.272	.14 (small)
		Label	9.90	.151	.237	.17 (small)
		Source × Label	-5.53	.559	.683	.10 (small)
Fixation Duration	AOI-1	Source	-51.39	.000	**.000** **	.43 (large)
		Label	-37.64	.006	**.017** *	.31 (medium)
		Source × Label	71.27	.000	**.000** **	.59 (large)
	AOI-2	Source	4.60	.038	**.069**	.14 (medium)
		Label	-2.80	.218	.343	.09 (small)
		Source × Label	1.75	.584	.642	.05 (small)
	AOI-3	Source	-0.46	.879	.879	.01 (small)
		Label	-1.61	.553	.642	.05 (small)
		Source × Label	2.85	.519	.642	.09 (small)
Saccade Count	AOI-1	Source	-5.13	.044	**.086**	.21 (medium)
		Label	-7.38	.013	**.047** *	.26 (medium)
		Source × Label	8.89	.047	**.086**	.36 (medium)
	AOI-2	Source	-4.35	.787	.787	.03 (small)
		Label	-7.82	.651	.716	.05 (small)
		Source × Label	43.77	.026	**.071**	.28 (medium)
	AOI-3	Source	-9.40	.223	.307	.12 (medium)
		Label	10.71	.160	.251	.14 (medium)
		Source × Label	-7.96	.487	.595	.10 (medium)
Saccade Length	AOI-1	Source	0.03	.124	.341	.17 (medium)
		Label	0.01	.531	.649	.08 (small)
		Source × Label	-0.07	.020	**.073**	.38 (medium)
	AOI-2	Source	0.00	.355	.558	.08 (small)
		Label	0.00	.181	.398	.12 (medium)
		Source × Label	0.00	.829	.829	.02 (small)
	AOI-3	Source	0.00	.276	.506	.08 (small)
		Label	0.00	.456	.627	.06 (small)
		Source × Label	0.00	.685	.754	.05 (small)
Pupil Diameter Fixation	AOI-1	Source	-0.41	.002	**.011** *	.35 (medium)
		Label	-0.33	.018	**.040** *	.29 (medium)
		Source × Label	0.50	.003	**.011** *	.43 (large)
	AOI-2	Source	0.00	.673	.823	.01 (small)
		Label	0.01	.445	.699	.02 (small)
		Source × Label	0.00	.898	.932	.00 (small)
	AOI-3	Source	0.00	.227	.416	.03 (small)
		Label	0.01	.531	.730	.01 (small)
		Source × Label	0.00	.932	.932	.00 (small)

Table 6.7: Results from the GEE analysis on the eye tracking data. (**p<.01, *p<.05, "p-value (corr)" denotes the result from GEE analysis with False Discovery Rate (FDR) correction.)

showed significant differences by source. Fixation count and saccade count showed significant interaction effects of source and label.

The post hoc comparisons shown in Fig 6.7 and Fig 6.8, participants demonstrated higher fixation counts ($p < .05$) and saccade counts ($p < .1$) in AOI-2 (main health information area) under the AI label condition, indicating that participants assessed the information focusing more on the content itself rather than the label when

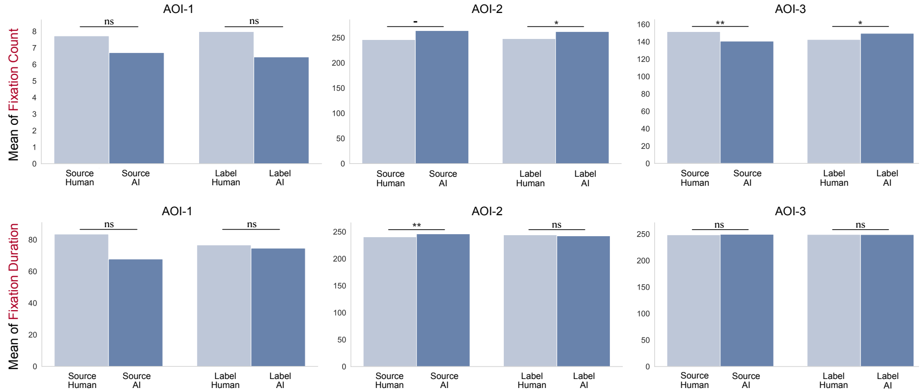


Figure 6.7: Posthoc pairwise comparison by Wilcoxon signed-rank test with False Discovery Rate (FDR) correction of fixation features (count and duration) in three AOIs. (**p<.01, *p<.05, -p<.10, "ns" is not significant).

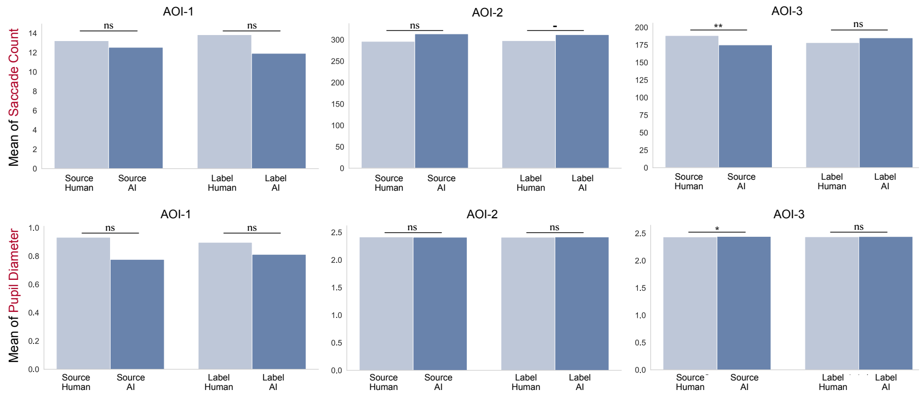


Figure 6.8: Posthoc pairwise comparison by Wilcoxon signed-rank test with False Discovery Rate (FDR) correction of saccade count and pupil diameter of fixation in three AOIs. (**p<.01, *p<.05, -p<.10, "ns" is not significant).

they were informed that the information is from AI. This implies that trust-related judgments in AI-labeled information were driven more by the actual content than the labeling of the source. Participants also showed significantly fewer fixation counts ($p < .05$) in AOI-3 (rating area) under the human label condition compared to the AI label condition. This suggests that human labels might inspire greater confidence, potentially influencing how users rate the trust score on the information. When information was actually sourced from LLM, participants showed higher fixation duration ($p < .01$) and counts ($p < .1$) in AOI-2, suggesting a more careful reading of AI-generated content. Conversely, human-sourced information led to higher fixation and saccade counts in AOI-3 ($p < .01$), indicating that LLM-sourced information might inspire greater confidence, potentially influencing how users rate the trust score, which aligns with the self-reported trust perceptions.

Analysis of physiological signals

Table 6.8 presents the results from GEE analysis on physiological data, shedding light on how physiological responses vary with different information sources and labeling.

Outcomes	Features	Conditions	Coefficient	P-value	Effect ($Std.\beta$)
ECG	BPM	Source (Human vs. LLM)	-0.58	.571	.07 (small)
		Label (Human vs. AI)	-1.10	.288	.13 (medium)
		Source × Label	1.38	.341	.17 (medium)
	RMSSD	Source (Human vs. LLM)	2.11	.435	.12 (medium)
		Label (Human vs. AI)	5.21	**.025** *	.29 (medium)
		Source × Label	-4.45	.179	.25 (medium)
	BPI	Source (Human vs. LLM)	8.88	.242	.12 (medium)
		Label (Human vs. AI)	10.43	.225	.14 (medium)
		Source × Label	-17.10	.153	.24 (medium)
EDA	SCL	Source (Human vs. LLM)	0.03	.949	.04 (small)
		Label (Human vs. AI)	-0.77	**.061**	.12 (medium)
		Source × Label	0.38	.414	.06 (small)
	SCR	Source (Human vs. LLM)	-0.56	.399	.05 (small)
		Label (Human vs. AI)	-0.92	**.082**	.08 (small)
		Source × Label	-0.98	.576	.08 (small)
Temperature	—	Source (Human vs. LLM)	0.46	**.022** *	.31 (medium)
		Label (Human vs. AI)	0.42	**.029** *	.28 (medium)
		Source × Label	-0.57	**.058**	.39 (medium)

Table 6.8: Results from the GEE analysis on the physiological signals. (**p<.01, *p<.05)

RMSSD, a feature derived from ECG data, was significantly higher for AI-labeled information compared to human-labeled information ($p = .025$). Higher RMSSD indicates greater heart rate variability (HRV), which is often associated with lower physiological arousal. This aligns with the gaze patterns that participants paid less attention to the labeling area (AOI-1) under 'AI' labels than 'Human' labels, as indicated by reduced fixation count and duration, saccade count, and pupil diameter (see Table 6.7, Fig 6.7 and Fig 6.8).

Skin temperature responses also varied significantly between human and AI labels ($p = .029$), as well as between human and LLM sources ($p = .022$). Higher skin temperature in response to AI labels and sources suggests participants may have experienced increased emotional arousal or stress when interacting with AI-associated content.

SCL ($p = .061$) and SCR ($p = .082$) average values did not exhibit statistically significant differences, as shown in Fig 6.9. This suggests that EDA components, at least within our study, were not discriminative of physiological arousal when users encountered human versus AI-generated health information.

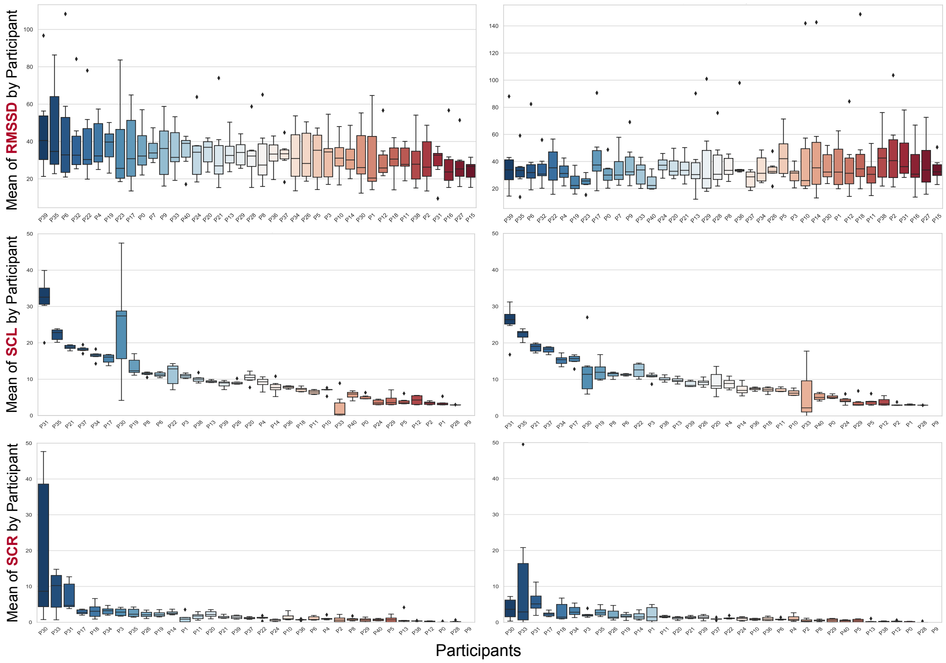


Figure 6.9: Pairwise comparison without correction on features of RMSSD and SCR per participant. **Left:** participants read the information labeled as from "Human Professionals" regardless of the source. **Right:** participants read the information labeled as from "AI" regardless of the source. Boxplots show individual participant-level distributions (median, IQR, outliers).

Correlation analysis

The Spearman correlation analysis [213] in Fig 6.10 and 6.11 revealed significant relationships between the self-reported trust score and various gaze and physiological features, indicating how participants' perceived trust in health information is linked to their behavioral and physiological responses.

Familiarity with the health question showed a strong positive correlation with trust in the information ($p < .01$). Among gaze features, fixation duration in AOI-1 (label area) positively correlated with the perceived trust score ($p < .01$), indicating that higher trust levels are associated with a longer focus on the labeling of information sources. Additionally, pupil diameter of fixation in AOI-1 ($p < .01$) also correlated positively with trust score. Fixation and saccade count in AOI-3 (rating area) were negatively correlated with trust, implying that participants who gave lower trust in the information exhibited more frequent saccadic movements in the rating area, likely reflecting efforts to evaluate or verify the information further.

No significant correlations were found between physiological features and trust levels. However, there were correlations observed among the physiological features

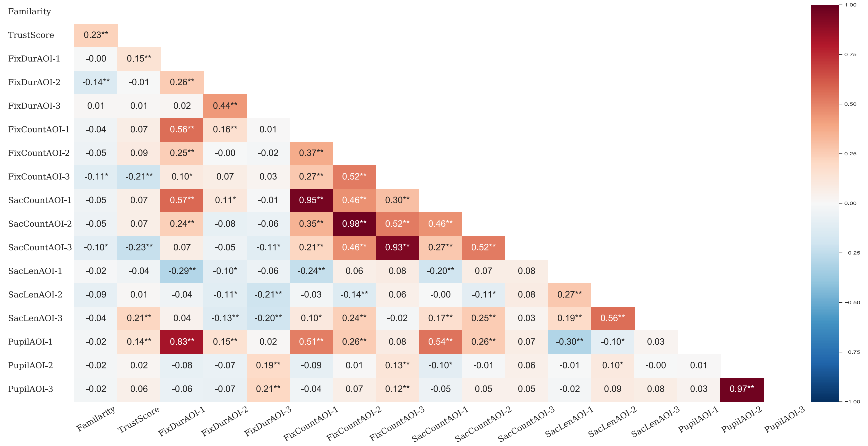


Figure 6.10: Spearman correlation with Bonferroni corrections between trust perceptions and the gaze features. (**p<.01, *p<.05). Note: "FixDurAOI-": fixation duration in AOI-; "FixCountAOI-": fixation count in AOI-; "SacCountAOI-": saccade count in AOI-; "SacLenAOI-": saccade lenth in AOI-; "PupilAOI-": pupil diameter of fixation in AOI-.

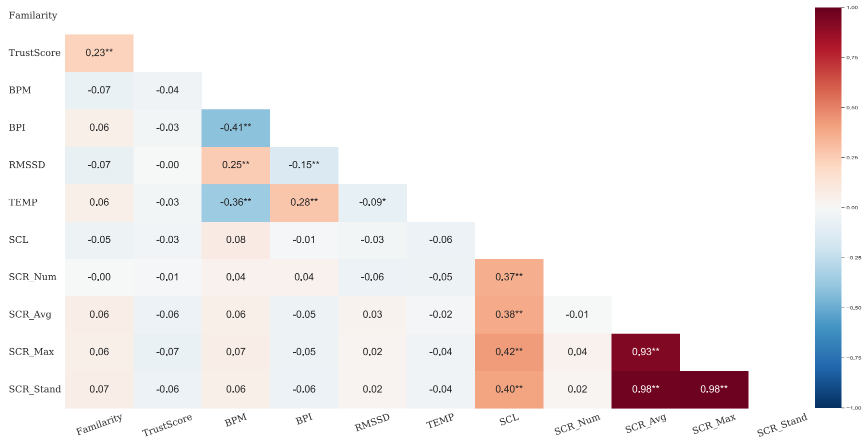


Figure 6.11: Correlation on variables. (**p<.01, *p<.05). Note: "SCR_Num": number of SCR; "SCR_Avg": average value of SCR; "SCR_Max": maximum value of SCR; "SCR_Stand": standard value of SCR.

themselves, such as BPM (Heartbeats), SCL, SCR, and skin temperature, though these did not directly link to trust.

Predictions using behavioral and physiological sensing

To explore trust perception (i.e., self-reported trust scores) through behavioral and physiological responses, we defined two tasks: 1) predicting participants' perceived trust score in health information and 2) classifying the source of the health information.

Models	Gaze Only		Physio Only		Gaze + Physio	
	MAE	R^2	MAE	R^2	MAE	R^2
SVR	.29	.06	.33	.06	.28	.10
Linear Regression	.25	.20	.31	.01	.24	.20
Ridge Regression	.24	.21	.31	.01	.24	.23
Random-Forest Regression	.23	.25	.25	.19	.20	.35
XGBoost	.24	.22	.28	.08	.23	.23

Table 6.9: Prediction on perceived trust score through regression using gaze and physiological features.

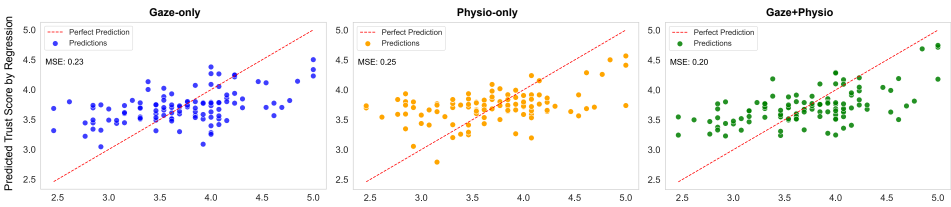


Figure 6.12: Prediction of perceived trust score using the Random-Forest Regression model on three different features set: Gaze-only, Physiology-only, Gaze+Physiology. Each dot represents one participant' s predicted vs. actual self-reported trust score, with the red dashed line indicating perfect prediction.

Features	Models	Trust Level		Source
		2-class (Acc / F1)	3-class (Acc / F1)	2-class (Acc / F1)
Gaze Only	LR	.65 / .62	.57 / .57	.62 / .55
	RF	.69 / .65	.57 / .57	.57 / .52
	SVM	.51 / .53	.43 / .42	.60 / .48
	MLP	.57 / .58	.32 / .32	.44 / .53
	GradientBoost	.72 / .66	.54 / .54	.52 / .52
	AdaBoost	.67 / .64	.58 / .58	***.65*** / .52
	XGBoost	.70 / .66	.54 / .54	.43 / .52
	Voting	***.73*** / .67	.54 / .54	.60 / .49
	Stacking	.70 / .66	***.59*** / .58	.49 / .55
	Bagging	.70 / .66	.57 / .57	.57 / .47
Gaze + Physio	LR	.65 / .62	.58 / .56	.58 / .54
	RF	.69 / .65	***.63*** / .63	.60 / .52
	SVM	.51 / .53	.43 / .43	.60 / .49
	MLP	.53 / .60	.48 / .47	.59 / .50
	GradientBoost	***.72*** / .68	.59 / .56	.53 / .53
	AdaBoost	.66 / .64	.54 / .54	***.65*** / .64
	XGBoost	.65 / .67	.57 / .57	.57 / .52
	Voting	.67 / .67	.59 / .58	.60 / .53
	Stacking	.69 / .66	.60 / .60	.48 / .52
	Bagging	.70 / .66	.61 / .61	.54 / .53

Table 6.10: Classification on binary trust level (high vs. low) and 3-class (high vs. medium vs. low), as well as on the source of information using gaze and physiological features.

For trust prediction, we first explored how regression models approximate perceived trust scores using regression models: Linear Regression (LR), Ridge Regression, SVM and Random Forest-based Regressions, and XGBoost. As shown in Table 6.9, the Random Forest regressor on the combined Gaze+Physio feature set achieved the lowest MSE of .20 and highest $R^2 = .35$ among the three feature sets, indicating the best performance. This highlights the value of combining gaze and physiological features for trust assessment. Fig 6.12 illustrates the regression results.

Next, we performed both binary (i.e., high vs. low) and three-class (i.e., high vs. medium vs. low) classification of trust levels based on participants' perceived trust scores. As shown in Table 6.10, the ensemble method (Voting model) achieved the highest accuracy (0.73) for binary classification using gaze-only features, while Random Forest achieved the highest accuracy (0.63) for three-class classification using combined gaze-physiological features. Notably, combining gaze and physiological features did not consistently improve performance across all models, for instance, the Gradient Boosting model achieved slightly lower accuracy (0.72) of binary classification when incorporating both feature sets compared to using gaze features alone. These results indicated that gaze features alone achieved higher classification accuracy for binary trust levels compared to combined gaze and physiological features. It suggests that gaze features may play a more prominent role in predicting trust levels than physiological responses, at least in the context of this study.

For the second task to classify the information source, combining gaze and physiological features yielded the best results. The AdaBoost model achieved the highest accuracy of 0.65 and F1 score of 0.64, indicating that physiological responses complement gaze features in distinguishing between human- and LLM-generated health information.

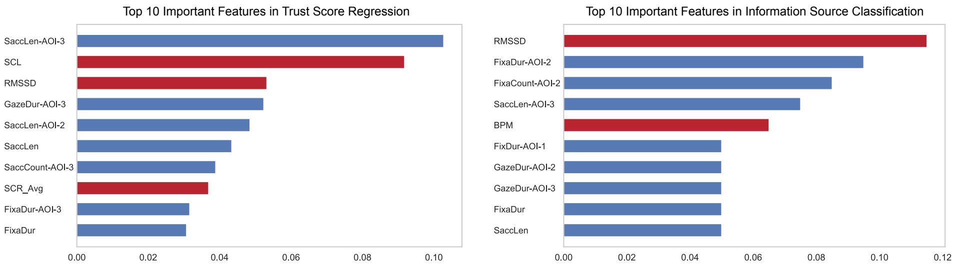


Figure 6.13: Top 10 important gaze (blue) and physiological (red) features in Random Forest regressor for predicting perceived trust score (**Left**) and in AdaBoost classifier for classifying the source of health information (**Right**).

Fig 6.13 presents the feature importance for the prediction tasks following the framework proposed by Lundberg and Lee [411] for better interpreting the model prediction. In summary, gaze features were effective for predicting trust perceptions, while combining gaze and physiological features could improve the classification of information sources. The robust performance of ensemble methods across both tasks highlighted their potential in developing tools to assess trust in health communication by leveraging gaze and physiological sensing.

6.5 Discussion

We conducted an online survey and a lab study in this work to investigate how users' trust responds to human versus AI-generated content, and in what ways trust in online health information may be modulated by including transparency labels as simple as "Human" versus "AI" labels on personal health information. Our findings showed that self-reported trust in digital health information is influenced by its actual source and disclosed labeling of the source. Further, the impact of these conditions were also evident at a behavioral and physiological level. Below, we discuss these aspects in detail.

6.5.1 Users may Prefer LLM-Sourced Health Information, but An AI Label Lowers their Trust

Both studies tested **(RQ1)** if the actual source, disclosed label, and type of information influence perceived trust in online personal health information. Our findings revealed that LLM-sourced content is trusted more than human-sourced content, regardless of labeling, whereas human professional labels are trusted more than AI labels. Trust however remained consistent across different information types (general, symptom, or treatment-related), suggesting that the source and labeling, rather than the type of information, are the primary determinants of perceived trust.

The observed difference in trust perception was evident in both self-reported trust scores (i.e., higher trust scores of LLM-generated information) and qualitative data, which suggests that participants have perceived subtle distinctions of information presentation styles in the LLM- versus human professionals-sourced information that provided cues for trust. The stronger effect observed in Study 2 (lab study with the within-subjects design) compared to Study 1 (survey study with the between-subjects design) further supports this, as the within-subjects design allowed participants to compare responses from both sources side by side. While we cannot conclusively determine the specific factors in information quality driving higher trust, our findings imply that LLM-generated content may convey an impression of clarity or objectivity that resonates more strongly with participants. Our observation that LLM-sourced information was trusted more than that from human professionals may reflect the advancements in LLMs like ChatGPT, which can produce structured and high-quality responses [283, 284]. Notably, GPT-4 generated responses have been found to be perceived as more human-like than actual human-authored content and other studies find that LLM-generated content is often indistinguishable from human-generated text [15]. This explanation (i.e., generally higher language quality of LLM-generated responses as a basis of trust) aligns with

Dalton et al.'s [29] proposal of emergent conversational information-seeking powered by LLMs, and is evident when assessing how LLMs are being used in the context of healthcare [33, 239, 240, 283, 284]. Furthermore, researchers also suggest that people prefer algorithms to humans in certain tasks and it could relate to individuals' machine heuristic (rule of thumb that machines are more secure and trustworthy than humans [366, 412]). In our studies, the qualitative analyses (Sec 6.3.3) further confirmed that participants attributed the higher trust in LLM-generated content to its efficiency, capacity to process extensive health data [185], and objective language style [32, 33]. This suggests that recent LLMs' (e.g., GPT-4 [196]) ability to deliver comprehensive and objective health information resonates with users, positioning them as reliable sources of health information.

Paradoxically, when health information was labeled as human, it was provided with higher trust scores than AI-labeled information, which is supported by Reis et al. [365], who found that people value human-labeled advice more when aware of AI's involvement, especially in the health context. This observation appears to generalize across various domains, whereby an AI label can diminish people's perceived quality, even if the AI source was initially deemed superior. This includes AI art [413], general communication [367], medical advice [377]. Even in clinical decision-making scenarios, people tend to prefer human decision-makers over AI, perceiving the latter as less dignified [414], further highlighting a deep-seated bias against AI involvement in sensitive health-related contexts. Moreover, Epstein et al. [415] found that not only the presence of a label, but also its wording, can significantly affect trust. For example, participants perceived content labeled as "AI-assisted" more favorably than "AI-generated", indicating that subtle linguistic framing influences users' willingness to trust. This suggests that beyond binary source disclosure, the design and language of labeling also play a critical role in shaping perception. Our findings (Sec 6.3.3) confirmed that participants expressed greater trust in human expertise, which they associate with verified knowledge, accountability, and human empathy. In contrast, they viewed the lack of consciousness, ethical judgment, and transparency in AI as diminishing their perceived trust. The perspective expressed by our participants aligns with De Freitas et al.'s [416] work about psychological factors affecting attitudes toward AI acceptance, which identifies opacity (lack of transparency or explainability) and emotionlessness (absence of empathy or moral understanding) as key factors driving user resistance to AI tools. Our respondents echoed these concerns by highlighting AI's lack of transparency and moral reasoning, especially in healthcare contexts, where trust is closely tied to perceived ethical awareness and human empathy. These reactions may also reflect a broader skepticism about machine consciousness [417]. Summarizing, although AI is recognized for its competence, our study showed that transparency in labeling human elements is crucial, emphasizing the need for transparent AI-powered systems [369] and authentic information [27, 418] to build trust, particularly when providing nuanced, personalized health advice [376, 377]. Clearly, there may be hidden dangers in adopting labeling approaches and relying solely on labeling as a trust mechanism (cf., [380]), as they might lead users to erroneously believe unlabeled content is accurate. This is related to the phenomenon what has been termed the "implied truth effect" [419].

6.5.2 Behavioral and Physiological Features Can Vary by Health Information Source and Label

Our results demonstrated that the effects of label and source are also evident at the behavioral and physiological level. Prior work has shown value in leveraging behavioral and psycho-physiological sensing across fake news detection in social media [384] and information-seeking tasks [382], where such signals are indicative of visual attention and information processing in these tasks. With respect to trust, Ajenaghughrure et al's. [56] review found that while psychophysiological levels of trust perceptions (e.g., arousal) can be detected (e.g., using EEG or ECG), how such responses behave during user interactions (in real-time) remains underexplored. In the context of our study, we first tested **(RQ2)** whether such signals vary during health information processing across human versus AI-sourced content, and essentially whether such signals can serve as a means of verifying and possibly even predicting trust (Sec. 6.3.1). We found that participants displayed distinct gaze patterns related to the source and labeling of the presented information. Specifically, we found that longer fixation duration, higher fixation counts, and larger pupil dilation were associated with information labeled as human-generated, suggesting a deeper cognitive engagement with this human-labeled information, suggestive of higher trust. Conversely, information labeled as AI-generated prompted more scanning behavior (i.e., reflected in increased saccadic movements and shorter fixation durations), indicative of increased verification processes. These results corroborate existing research from others (e.g., Just et al. [420] and Rayner et al. [421]) who likewise found that gaze patterns, especially the fixation and saccade behaviors, are indicative of cognitive processing and information verification relevant to trust assessment and dynamics.

For the peripheral physiological signals, we found significant differences in features such as RMSSD and skin temperature when users encountered labeled health information. No differences were found in skin conductance (SCL and SCR) measurements. It is worth speculating what this means: these indicators aligned with users' self-reports, whereby health information labeled as from AI elicited higher HRV (i.e., RMSSD) than the label of human professionals. Higher HRV is typically associated with lower physiological arousal, possibly reflecting less cognitive processing or more relaxed state. This interpretation is consistent with the meta-analysis by Kim et al. [391], which found that HRV reliably decreases under stress or increased cognitive demands, and increases under lower arousal or more comfortable conditions. Indeed, HRV is one of the most commonly used psychophysiological indicators in trust research [56], able to detect subtle variations in user state during human-computer interaction. Although Ajenaghughrure et al. caution that trust classification using physiological signals remains an open research challenge. Furthermore, the pattern of reduced physiological arousal in response to AI-labeled information aligns with the gaze data in our study, which suggested less attentional engagement (e.g., shorter fixations, fewer regressions) with AI-labeled content compared to human-labeled information. These findings suggest that participants may have processed AI-labeled health information with lower cognitive and emotional investment. Similarly, higher skin temperature levels were observed with both AI-labeled and LLM-sourced information, suggesting lower emotional arousal and stress levels, aligning with participants' psychological interpretation of trust [61]. I.e., participants gave higher trust scores to the LLM-sourced information compared to human-sourced, and showed less physiological arousal with the AI labels than human labels. Taken together, these physiological signals could serve as a useful means to corroborate how users react and feel toward content perceived to be sourced from humans versus AI systems, while providing an additional layer of information about information processing and associated affect.

6.5.3 Towards Disclosure-Aware Interfaces in AI-Powered Health Systems

Our research touches upon recent efforts in Human-Computer Interaction to draw on physiological signals when designing or evaluating interactive systems [59], for example through either cognitive load estimation using physiological and gaze data [381]. These efforts are directed towards developing bias-aware systems by utilizing physiological and interaction data [422], or more specifically, on predicting trust using psychophysiological measures [393]. Hence, in Study 2, we explored the utility of behavioral and physiological signals (e.g., eye movements) to predict self-reported trust scores and the ability to discern the actual source of information. In a user-independent setting, we achieved 0.35 R^2 in a regression task predicting perceived trust scores, and 73% accuracy if we bin the trust scores (high vs. low) and 63% accuracy if we bin the trust scores (high vs. medium vs. low). Furthermore, we reached 65% accuracy in classifying the information source. While these accuracies are modest and need independent replication, they provided a hint towards a future of adaptive health information systems whereby a combination of behavioral and physiological insights may be used for trust assessment beyond self-report, as well as monitoring "depth" of attention and processing of information. Such sensors would open the possibility of not only verifying trust perceptions but also, based on sensing data, automatically inferring if and when to apply transparency labels on user interfaces. I.e., given the critical role of UX in ensuring responsible AI in designing and deploying AI systems [423], information systems can benefit from becoming "disclosure-aware", whereby the label is shown if sensed behavioral and physiological data requires it. Trust can hereby already be scured at the interface level when presenting health information [50]. Such disclosure-aware interfaces could be a step toward reducing over- and under-reliance on AI-generated content (cf. [49, 51, 424]), particularly when it is known that source labels can influence trust. For example, the interfaces could emphasize the AI origin of the information when behavioral or physiological signals suggest over-reliance (e.g., accepting information without adequate scrutiny), or withhold unnecessary cues when signs of under-reliance are detected (e.g., premature skepticism toward accurate AI-generated information).

6.5.4 Limitations and Future Work

Our study had several limitations that should be considered when interpreting findings.

First, we are cautious regarding the predictive results of physiological and behavioral signals. While we have observed spikes in physiological activity and eye movement markers, we advise against simplistic interpretations that human trust evaluations (a complex and subjective phenomenon [254,256,371]) in health information can be fully explained, or even be replaced, by such markers. Furthermore, physiological signals can be indicative of multiple phenomena from attention to physical excitation to underlying medical conditions [425]. Without careful contextualization, these signals could be misinterpreted as significant in scenarios where they merely represent contextual noise. Future work could incorporate additional sensor modalities (e.g., fNIRS [422] or EEG [426]) to provide more robust measures of cognitive activities for trust verification.

Second, the controlled lab environment may have influenced participants' responses, as being observed might heighten scrutiny of AI-labeled information, potentially amplified by societal caution toward AI. However, such an a "mere observer effect", are likely just typical for controlled psychological experimental conditions, where participant awareness of observation can subtly affect behavior [427]. While these settings are valuable for minimizing external confounders and ensuring reliable comparisons across conditions, future studies should nevertheless validate these findings in real-world environments to account for potential differences in naturalistic behaviors.

Third, our study did not assess participants' actual actions or decision-making following the health advice provided. One of the core aspect of trust is its propensity to steer decision-making in the context of uncertainty, even when this decision-making involves taking risk (see Chapter 1 in this thesis). Real-world decision-making, such as whether participants follow AI- or human-sourced advice in practice may thus provide a more robust measure of trust. In future work, we therefore plan to employ action-oriented trust paradigms through longitudinal experimental designs, such as Ecological Momentary Assessment (EMA) [428]), which can observe participants' decisions and behaviors in real-world contexts and provide a more comprehensive understanding of trust dynamics. For the applicability of these results in developing real-world solutions (e.g., using web-based gaze trackers [429] or rPPG, a technique that estimates heart rate and other physiological signals from facial videos [430]), ethical concerns over accountability [372], data privacy breaches, and potential over-reliance on AI [37] must subsequently also be considered. hence, well-developed frameworks are required whereby behavioral and physiological sensing involves continuous consent based on robust on-device security and privacy controls, compliant with legal regulations (e.g., European AI Act [431]).

Fourth, in both survey and lab experiment, measurements were taken at a single point in time. While practical, this approach may not fully captures dynamic nature of information-processing and trust [261]. Include multiple time points or trials per condition will reduce individual variability, hereby improving robustness, in addition to establishing trust dynamics over time. Longitudinal designs would provide knowledge of how trust in health information is influenced which may generalize better to how LLMs are being used in the real-life digital health context. Additionally, while all LLM-generated responses were reviewed for accuracy and alignment with health guidelines, we did not explicitly screen for stylistic differences such as tone and uniformity, which might also influence trust perceptions. Future work should systematically evaluate the role of stylistic linguistic features, such as clarity, tone, and perceived quality, with regard to the trust in AI.

Lastly, our participant sample (notably WEIRD [432]) across both studies was not representative of the general population, further limiting generalization. This is particularly relevant for groups with varying levels of AI literacy or differing baseline trust in technology. Acknowledging this limitation helps specify to whom these findings most apply. Nevertheless, our study provides a key initial step toward understanding the impact of source and labeling in online health information. Future expansion to include participants from varied demographics can enhance our understanding of how trust in health information is perceived across different groups.

Overall, this study highlighted how AI labeling influences trust in health information in controlled settings using a homogeneous sample. These findings underscored the need for future research using real-world, longitudinal designs for capturing trust-related behaviors over time. Addressing these gaps will be essential to inform the design of AI health systems that are both trustworthy and inclusive.

6.6 Conclusion

Through a mixed-methods crowdsourcing survey (N=142) and within-subjects lab study (N=40), we found that AI-generated health information is trusted more than content sourced by human professionals, regardless of labeling, while human labels are trusted more than AI labels. Furthermore, we found that trust perceptions in personal health information are not only influenced by the source and label but also vary at behavioral and physiological levels. Our work highlighted the importance of considering AI transparency labels when measuring trust in online health information, and in developing techniques for verifying subjective trust perceptions and automatically inferring if and when to apply transparency labels based on sensed behavioral and physiological data. As such, we invite future research on understanding and designing for the physiology of online human-AI interactions, within and beyond AI-powered health information systems.

7

General Discussion

7.1 Answering the Research Questions

This thesis explored two interconnected themes in LLM-powered digital health: (1) aligning LLMs with domain expertise for controllable and explainable digital health behavioral intervention; and (2) investigating how trust in LLM-powered health communication is influenced by user interface and source transparency, using mixed methods including behavioral and physiological sensing. Across multiple empirical studies, we examined the capabilities and limitations of LLMs in health intervention and communication, addressing the research questions. This chapter provides a general discussion of the findings for each research question, as well as the implications of the research presented in the thesis.

7.1.1 Aligning LLMs with Domain Expertise for Health Behavioral Intervention

RQ 1: How well do LLMs perform in generating reflective utterances in Motivational Interviewing (MI) sessions? (Chapter 2)

In Chapter 2, we systematically assessed LLM's performance across multiple language models and prompting strategies to evaluate its ability to generate reflective utterances in MI sessions. The primary objective was to determine whether LLMs can generate contextually appropriate, specific, natural, and engaging reflections, aligning with the core principles of MI. The evaluation combined automatic metrics (e.g., BERTScore and BLEURT) with human assessments, providing a comprehensive evaluation and impression of LLM-generated reflections in MI.

Findings from human assessments in Chapter 2 suggested that LLMs, particularly GPT-4, can generate reflective utterances that are perceived as comparable to human-authored reflections in terms of appropriateness and specificity. GPT-4 consistently outperformed Llama-2 and BLOOM, producing reflections better aligned with MI principles, suggesting its stronger capacity to process longer conversational context and maintain relevance to client statements. Additionally, context length played a significant role in reflection quality. When models were provided with a longer conversation history (five turns instead of one turn), their reflections were rated as more appropriate and contextually aligned. This suggests that LLMs benefit from extended conversational context, enabling them to generate reflections that are more coherent and connected to previous interactions. However, engagement scores slightly declined with longer contexts, indicating that while LLMs could maintain contextual relevance, their ability to produce engaging and dynamic responses may be constrained when managing extensive conversation histories. This trade-off may echo broader patterns observed in SSAG in Chapter 4, where certain response types, such as therapeutic questions and advice, were retrieved and paraphrased from expert-crafted utterances to ensure safety and alignment. Although not directly compared with the study in Chapter 2, one might expect similar effects on engagement: longer context with pre-scripted or paraphrased responses, while appropriate and relevant, may give less emotionally resonant to users. This raises interesting design questions around balancing contextual grounding with emotional engagement in multi-turn psychotherapeutic conversations.

While LLM-generated reflections were often rated highly in appropriateness and specificity, they fell short in capturing the deeper emotional nuances of human therapists. Our manual content analysis of generated reflections revealed that LLMs sometimes made implicit assumptions about client statements rather than eliciting deeper self-reflection, a key function of MI. This indicated that while LLMs can mimic MI techniques to some extent, they lack the intuitive adaptability and emotional intelligence of trained human therapists.

Overall, LLMs, particularly GPT-4, demonstrated strong potential in generating MI-consistent reflections, especially when provided with extended conversational context. However, their performance remained context-dependent, and they still exhibited limitations in emotional depth, engagement, and flexibility in adapting to dynamic therapeutic interactions. These findings highlighted the feasibility of using LLMs as supportive tools for MI-based interventions while emphasizing the need for further refinements in their ability to generate emotionally resonant and adaptive reflections.

RQ 2a: Do expert-crafted dialogue scripts remain essential for chatbot-delivered psychotherapy in the era of LLMs? (Chapter 4)

LLMs demonstrated strong potential and capability in generating specific dialogue utterances called "Reflections" in MI (Chapter 2). Therefore, we continuously investigated how LLMs aligned with domain expertise perform in generating psychotherapeutic dialogues for health behavior intervention in Chapter 3 and 4.

RQ 2a explored Script-Aligned Generation (SAG) as a key method to align LLMs with domain expertise (i.e., expert-crafted dialogue scripts) for delivering psychotherapy. To thoroughly examine it, we divided RQ 2a into two subcomponents: RQ 2a (1) evaluated whether LLMs aligned with expert-crafted dialogue scripts outperform rule-based chatbots and pure, unaligned LLMs; and RQ 2a (2) compared two approaches for script alignment: fine-tuning and prompting, to determine which is more effective and efficient. Together, these questions help assess both the necessity of expert scripts in the LLM era and the most efficient method for aligning LLMs.

The findings for RQ 2a (1) showed that SAG-aligned chatbots significantly outperform both rule-based chatbots and unaligned pure LLMs across key therapeutic metrics. While rule-based chatbots ensured strict safety and adherence, they lacked adaptability. Pure LLMs, although more fluent and engaging, often failed to follow therapeutic structures and goals. In contrast, SAG combined the structure of expert-crafted scripts with the adaptability of LLMs, enabling conversational interactions that are both therapeutically grounded and conversationally natural. This highlighted that expert-crafted dialogue scripts remain essential, not as rigid templates, but as flexible scaffolds for LLM-powered psychotherapy. To answer RQ 2a (2), prompting was found to be a more effective and efficient alignment method than fine-tuning. Prompted SAG LLMs better preserved therapeutic structure and offered better conversational flexibility and engagement, while requiring fewer resources than fine-tuning. This makes prompting a scalable and adaptable approach for aligning LLMs with expert-designed therapeutic content in real-world applications.

RQ 2b: Can psychotherapy chatbots using script-strategy aligned generation (SSAG) achieve comparable conversational quality and therapeutic effectiveness to those using script-aligned generation (SAG)? (Chapter 3 & 4)

This question examines whether Script-Strategy Aligned Generation (SSAG), an extension of script-aligned generation (SAG), can reduce expert scripting demands while maintaining therapeutic structures. Unlike SAG, which requires full expert-crafted dialogues, SSAG combines partial expert-scripted dialogues (e.g., therapeutic topics and key questions) with MI strategy prediction to guide dialogue generation. This approach integrates both script and strategy alignment, echoing prior work that highlights the value of therapeutic expertise in instructing psychotherapy [158].

To assess SSAG ' s feasibility, we first validated LLMs' ability to predict relevant MI strategies for the next conversational turn (Step 1 of SSAG). Results on two benchmark datasets (AnnoMI and BiMISC, Chapter 3) showed that models like GPT-4 can effectively identify appropriate strategies, an essential step for generating coherent and goal-directed therapeutic responses subsequently. Automatic evaluations (i.e., BLEU, ROUGE, METEOR, BERTScore and Entropy) further confirmed that strategy-aligned prompting can improve the generated response's alignment with MI principles. Moreover, open-source models (e.g., Qwen-14B, Llama-2-13B) also performed competitively with GPT-4, highlighting the scalability of SSAG using publicly available models.

In Study 2 of Chapter 4, we compared SSAG with SAG (prompting-based). Findings showed that SSAG achieved comparable performance across evaluation criteria such as empathy, engagement, MI adherence, and relevance, suggesting that while expert-authored scripts provide important therapeutic structure and fidelity, they do not need to be used in full. By anchoring each generated utterance to a predicted MI strategy, SSAG preserved structure and therapeutic quality while enabling more adaptive and flexible dialogue generation. Participants described SSAG chatbots as "structured but flexible", reflecting the benefit of strategy-level alignment without full reliance on the expert-crafted dialogue scripts. In sum, SSAG offered a promising middle ground, maintaining therapeutic effectiveness with reduced expert input, enabling scalable, explainable, and controllable development of psychotherapy chatbots.

RQ 2c: To what extent can SSAG reduce the reliance on expert-scripted content for developing the psychotherapy chatbots? (Chapter 3 & 4)

The work in Chapter 4 demonstrated that Script-Strategy Aligned Generation (SSAG) can reduce reliance on fully expert-crafted dialogue scripts while preserving therapeutic quality. Traditional rule-based chatbots and the LLM-powered chatbots employing SAG (Study 1 of Chapter 4) require extensive domain expert effort to author structured dialogues, limiting their scalability. SSAG addressed this by separating dialogue generation from full-scripted control: experts define high-level therapeutic strategies, topics, and core questions, while LLMs generate responses guided by the LLM-predicted MI strategies for the next conversational turn. Empirical results showed that SSAG uses less than 40% of expert-authored content, achieving comparable performance to the aligned LLM chatbots that use the full expert-crafted dialogues (i.e., SAG). This demonstrates a substantial reduction in manual scripting workload. The two-step SSAG process enables LLMs to produce adaptive, context-aware dialogues while remaining anchored to domain principles.

From a broader human-AI collaboration perspective, SSAG enables structured division of labor, LLMs handle generation within expert boundaries, while preserving explainability and controllability. Through this lens, this work provides a practical approach for expert-AI co-development and offers a promising step toward building cost-effective, expert-informed and socially impactful AI technologies at scale that advance mental health support, particularly in areas where expert resources are limited.

7.1.2 Understanding Trust Perception in LLM-Powered Health Communication

RQ 3a: How does people's perceived trust in health information vary across different search agents (e.g., Google vs. ChatGPT) and dissemination interfaces (text-based, speech-based, or embodied)? (Chapter 5)

RQ 3b: How does inherent trust in these agents and interfaces influence the trust perception in health information? (Chapter 5)

Trust is a critical determinant of the adoption of health information from digital systems [24], especially from AI or LLM. Our research in Chapter 5 revealed that perceived trust in health information significantly differed and was influenced by both the search agent and the dissemination interface.

Study 1 found that participants trusted health information from ChatGPT more than Google. Interviews revealed that this higher trust was driven by LLM's interactive style, which made users feel heard and understood, as well as its personalized and rapid responses, which enhanced perceived relevance and expertise of the system. Additionally, ChatGPT's user-friendly interface was seen as intuitive to engage with, further increasing its credibility. LLM-powered conversational agents [29], such as ChatGPT, generate synthesized responses rather than directing users to external sources. While this approach enhances accessibility and engagement, it also introduces trust-related challenges, including lack of source attribution, potential biases, perceived AI involvement [363, 365], and authenticity of AI-generated content [418]. These findings underscored the need for careful design in LLM-powered health systems to balance convenience with source transparency and reliability. In contrast, trust in Google's health information was not primarily based on interaction design or response style but rather on users' longstanding familiarity with the search engine. Trust in Google's health information was strongly correlated with inherent trust in Google itself, which appeared to be inherited from the brand or familiar searching behavior. It suggested that historical use and perceived authority play a strong role in shaping user confidence in searching with a traditional search engine such as Google. No such correlation was found for ChatGPT, indicating that users may still be developing trust in LLM-powered search and distinguish between trust in system and trust in information it provides. These differences may also be attributed to the fundamental contrast between retrieval-based and generative-based search modes. Traditional search engines like Google present ranked results from multiple sources, encouraging active information-seeking, verification, and cross-checking.

Beyond the interactive mode how people search information, the interface through which information is disseminated, through text-based, speech-based, or embodied modality, also plays a crucial role in shaping trust [30]. Study 2 of Chapter 5 explored how dissemination interfaces affect trust perceptions. Participants favored text-based interfaces due to their familiarity and ease of verifying information, particularly in high-stakes health contexts. While speech-based and embodied interfaces provided more engaging and natural interactions, they did not inspire the same level of trust. The preference for text-based interfaces stemmed from their ability to support deeper cognitive processing and facilitate cross-referencing, which is essential when evaluating health advice. Correlation analysis further confirmed that inherent trust in the interface directly influenced trust in the information it delivered. Findings from interviews showed that speech-based interfaces increase social presence and immediacy, which can enhance engagement but also introduced challenges such as cognitive overload, difficulty in recalling verbal information, and a lack of control over the information presented. Similarly, embodied interfaces, such as robotic assistants, introduced non-verbal communication cues that amplify the sense of human-like interaction. However, trust in embodied interfaces remained highly context-dependent; while embodiment can enhance credibility in general AI interactions, it may raise concerns regarding authenticity, privacy, and potential biases associated with anthropomorphic AI representations, especially in health domains, where accuracy and transparency are paramount.

These findings underscored the importance of not only the content delivered by LLMs but also the design of LLM-powered health applications. Trust in search agents and dissemination interfaces directly influenced how users interpret and evaluate the health information provided. Designing future LLM-powered systems for health information seeking will require an understanding of how to balance familiarity, information credibility, and interactivity across both agent and interface dimensions, ensuring that conversational systems incorporate verifiable information sources and transparency mechanisms to maintain credibility in health communication.

RQ 3c: What factors contribute to trust perception across search agents and dissemination interfaces? (Chapter 5)

While RQ3a and 3b address what participants trust, this sub-question (RQ3c) focuses on “how”. From qualitative findings in Chapter 5, several key factors emerged that affect perceived trust in health information, highlighting the interplay between search agents, dissemination interfaces, and trust perceptions.

A central factor is source transparency. Participants trusted information more when the source was verifiable and explicitly cited. Google benefited from providing multiple sources and links for cross-checking, whereas ChatGPT's lack of references raised skepticism, leading some participants to doubt the credibility of its responses. Participants suggested that direct source attribution could enhance trust in LLM-powered tools, making them more reliable for health-related inquiries. Another key factor was prior experience and familiarity with AI systems. Individuals with previous exposure to ChatGPT or similar LLMs tended to express higher trust in their outputs. Similarly, participants more accustomed to text-based interfaces reported greater trust, whereas speech-based and embodied modalities required longer adjustment periods to establish trust. Search autonomy (i.e., perceived user control), particularly the ability to independently verify and choose among information, emerged as another trust factor. Google's open-ended retrieval model fostered this sense of autonomy, whereas ChatGPT's one-shot generative responses, though efficient, were viewed as opaque to audit, especially in health contexts. Furthermore, cognitive processing mode affects how users engage with information across different interfaces. Text-based dissemination supported deeper information scrutiny and facilitated deeper cognitive processing, while speech and embodied interfaces introduced higher cognitive demands, leading to surface-level trust judgments based on interaction perceptions rather than factual accuracy of information.

In addition to these factors, machine intelligence and personalization brought by LLM also play a significant role in affecting trust in AI-driven health communication. Participants trusted agents that provided contextually relevant, well-structured responses, while generic or inconsistent replies reduced credibility. Personalization can further influence trust in mixed ways, depending on user preference. Some participants found tailored responses based on past interactions helpful and confidence-enhancing. Others raised privacy concerns and feared that excessive personalization could introduce bias or reduce exposure to diverse perspectives. Trust was highest when personalization was transparent and users could control how the system adapted its responses.

Taken together, these factors underscore that trust in LLM-driven health information seeking is influenced not only by what is presented but also by whom and how the information is presented, pointing to the need for transparent, user-adaptive, and verifiable systems and interfaces in future health technologies.

RQ 4a: How do the actual source, disclosed label, and type of health information influence perceived trust? (Chapter 6)

Building on the prior finding in Chapter 5 that trust is shaped not only by what information is presented but also by who presents it and how, the work in Chapter 6 delved deeper into the roles of information source, disclosed label, and health information type in shaping trust perceptions. The findings revealed a complex interplay between these factors in shaping trust perceptions. Across both the online survey and laboratory experiments in Chapter 6, we observed that LLM-generated health content was generally trusted more than human-authored content, regardless of the disclosed labeling, suggesting that users may perceive LLM-generated information itself as reliable and consistent. However, trust shifted when explicit labels were introduced, with information labeled as human-generated being trusted more than AI-labeled content, regardless of the actual source. This discrepancy highlighted a cognitive bias in trust calibration, where users rely on labeling cues even when they do not necessarily reflect perceived content quality. Notably, trust remained stable across health information types (general, symptom, and treatment-related), suggesting that the influence of source and label transparency operates consistently across various health contexts. This implied that users may form generalized trust heuristics based on AI involvement rather than evaluating credibility based on specific information, as investigated in prior work [365]. This finding was particularly relevant for the design of AI-driven healthcare systems, as they underscored the need for transparency mechanisms to align user expectations with actual content reliability.

RQ 4b: Can behavioral and physiological signals provide insights into trust perceptions toward human- and AI-generated health information? (Chapter 6)

To gain deeper insights into how trust manifests beyond self-reported trust level, we analyzed behavioral and physiological data from the lab study. Participants exhibited distinct gaze patterns and physiological responses when interacting with LLM- versus human-generated information and AI- versus human-generated labels. Gaze patterns, including fixation count, fixation duration, saccade count, and pupil diameter, revealed that participants paid less attention to the labeling area when information was AI-labeled, potentially indicating lower cognitive effort or reduced scrutiny when AI was disclosed as the source. This suggested that users may apply a different cognitive processing strategy when engaging with AI-labeled information, relying more on content evaluation rather than label-based trust heuristics.

Physiological responses further reinforced these patterns. Higher RMSSD (indicative of increased heart rate variability and lower arousal) was observed when participants interacted with AI-labeled information, indicating lower physiological arousal. Conversely, human-labeled content elicited stronger physiological and behavioral responses, indicating greater cognitive investment in evaluating human-labeled information.

To further validate the potential of behavioral and physiological sensing in detecting trust, we trained machine learning models to classify participants' trust levels based on collected physiological and gaze data, using self-reported trust as the ground truth. Our results demonstrated that gaze data alone achieved 73% accuracy in distinguishing between high and low trust levels. Besides, a combination of gaze and physiological data reached 65% accuracy in identifying the actual source of information (LLM vs. human). These findings suggested that trust is not only reflected in explicit self-reports but could also be inferred from implicit behavioral and physiological indicators, offering new directions for designing AI-powered health interfaces that adapt to users' trust perceptions in real time.

By integrating explicit self-reported trust assessments, implicit behavioral and physiological sensing, our study provides a complementary understanding of how trust in health information is influenced by both information source and label. Results highlighted complex interplay between cognitive and physiological processes in trust formation and underscored importance of transparent AI disclosures in health-related decision-making. Moreover, the ability to predict trust by behavioral and physiological data suggested promising applications for real-time trust-aware AI systems or user interfaces with regulatory frameworks such as GDPR [433] or European AI Act [431], which could help mitigate misinformation and enhance user engagement with credible health content.

7.2 Implications

This thesis provides critical insights into the role of LLMs in digital health, with a dual focus on aligning them with domain expertise for health interventions and understanding people's trust formation in LLM-driven health information seeking. The findings have broad implications for advancing LLM alignment techniques to enhance explainability and therapeutic adherence, rethinking trust mechanisms in LLM-generated health information, and addressing the ethical and regulatory challenges of deploying human-centered AI in digital health. These insights underscore the need for AI systems that not only leverage technical advancements but also prioritize transparency, trustworthiness, and ethical responsibility in healthcare applications.

Aligning LLM with Domain Expertise for Therapeutic-Adherent Health Intervention

The findings of this thesis underscore the importance of aligning large language models (LLMs) with expert therapeutic expertise, such as motivational interviewing (MI) strategies and expert-crafted dialogue scripts, to ensure their effective, reliable, and responsible use in psychotherapy for health interventions. While LLMs exhibited strong capabilities in generating fluent and contextually appropriate responses, their lack of inherent adherence to evidence-based therapeutic principles and structures, such as MI skill codes and CBT structures, limits their reliability and therapeutic efficacy in sensitive psychological counseling contexts. This research demonstrates that without proper alignment, LLM-generated dialogues may deviate from evidence-based therapeutic principles, reducing their effectiveness. To address this, this thesis demonstrates that integrating explicit reasoning about therapeutic strategies into LLMs improves both controllability and expandability. This alignment not only enhances the quality of LLM-driven therapeutic conversations but also contributes to broader goals of explainable AI (XAI) research [434]. Making LLMs' generation transparent is shown to be a critical step toward explainability, and ensuring controllable deployment of AI in mental health and other high-stakes domains.

To achieve a more controllable and explainable integration of LLMs into psychotherapy, this thesis explores an expertise-driven alignment, where LLMs are first guided to predict MI strategies before generating responses aligned to the expert-crafted dialogue scripts. This approach significantly improves therapeutic adherence, response coherence, and explainability, ensuring that LLM-generated dialogues follow recognized therapeutic techniques rather than relying solely on statistical text generation [435]. Both computational and empirical findings validated that SSAG enhanced the controllability and explainability of LLM-driven interventions, making them more consistent with domain expertise.

Beyond MI, the implications of this work extend to broader psychotherapeutic applications. Many counseling frameworks, such as cognitive behavioral therapy (CBT) [17] or dialectical behavior therapy (DBT) [436], involve interventions that require domain expertise for effective and efficient implementation. The alignment approach developed in this thesis provides a foundation for adapting LLMs to other therapeutic frameworks, ensuring that LLM-driven health interventions remain therapeutically valid and ethically sound. Furthermore, this research highlights the critical balance between therapeutic adherence and dialogue empathy in LLM-generated dialogues for health behavior intervention. While expertise-driven alignment (SSAG) improves adherence to therapeutic principles (i.e., MI strategies and CBT structures), it remains challenging for LLMs to fully capture the emotional nuances, adaptive responses, and long-term contextual awareness exhibited by human therapists. This calls for future advancements in AI personalization, multi-turn memory retention, and contextual adaptation, ensuring that LLM-powered psychotherapy applications provide not only structured therapeutic responses but also genuine empathetic engagement [437].

By aligning LLMs with expert-driven strategies and dialogue scripts, this thesis could contribute to the development of explainable, controllable, and therapeutically effective LLM systems for mental health support and behavioral intervention. The work in this thesis paves the way for AI-assisted digital psychotherapy that functions as a complementary tool for therapists, digital coaching aids for patients, and scalable mental health support systems for broader populations, ensuring that LLM-powered health interventions are both therapeutically effective and ethically responsible.

Rethinking the Trust Perception in LLM-driven Health Communication: The Role of Interaction Type, Dissemination Interface and Source Transparency

As AI-powered health communication becomes increasingly prevalent, trust perception plays a critical role in determining the effectiveness and adoption of LLM-generated health information. The findings of this thesis have significant implications for the design of AI-driven health communication systems, the interaction modes for health information-seeking, the development of transparency and cognitive mechanisms for trust calibration in digital health interactions.

One key implication is the need to rethink the design of AI-powered applications, including both models and user interfaces, to promote informed trust rather than blind reliance. Findings from comparative studies between traditional search engines (e.g., Google) and LLM-powered conversational agents (e.g., ChatGPT) revealed that users generally exhibit higher trust in AI-generated health content due to its coherent, direct, and seemingly authoritative presentation. However, this heightened trust does not necessarily stem from increased accuracy but rather from the interactive, structured, and personalized nature of conversational AI, which fosters a perception of expertise and credibility. Moreover, studies examining the effect of dissemination interfaces (text-based, speech-based, and embodied) revealed significant trust variations based on interaction style. While text-based interactions remain the most trusted due to their familiarity and ease of verification, speech-based and embodied interactions introduce higher social presence, influencing users' cognitive and affective engagement with information retrieved. The findings suggest that users associate richer interaction modalities with higher engagement, yet in high-stakes health scenarios, they still prioritize transparency and verification over interactivity. To mitigate the risk of automation bias, LLM-driven health interfaces should integrate features that encourage critical evaluation, such as interactive explanation displays, source attribution, and mechanisms that enable users to compare information from multiple sources. Future AI-powered health systems should be designed to balance ease of access with mechanisms that encourage users to assess the credibility of AI-generated information actively without over-reliance on AI.

Transparency mechanisms, particularly source disclosure (AI-generated vs. human-generated labels), play a critical role in modulating trust perception, but their effects are not always straightforward. This thesis found a paradoxical effect in AI labeling, while AI-generated content is generally trusted more than human-generated content, explicitly labeling information as AI-generated reduces perceived trust compared to human-labeled content, regardless of its actual source. This echoes the work from Reis et al. [365], showing that medical advice labeled as AI was perceived as less reliable and empathetic than advice labeled as human-generated, despite being identical in content. This suggests that users hold implicit biases associating human expertise with credibility, even when AI-generated content may be more objectively accurate and well-structured. This underscores the need for more adaptive transparency strategies rather than relying solely on rigid labeling practices. Future AI health applications can explore context-sensitive transparency, where explanations are tailored based on user familiarity, prior experience with AI, and the complexity of the health information being presented. For example, novice users may benefit from more explicit AI disclosures, while experienced users may prefer minimal but well-integrated credibility indicators.

The findings of this thesis also highlight that trust in AI-generated information is not solely determined by content quality but is influenced by cognitive processing, inherent trust, and user interfaces. Users with longstanding familiarity with traditional search engines (e.g., Google) exhibited higher inherent trust in the information retrieved from them, while trust in newer AI-powered search tools (e.g., ChatGPT) remained less stable and continues to develop over time. This suggests that trust perception is influenced not only by the accuracy or reliability of content but also by historical experiences, perceived source credibility, and interaction familiarity. Such biases can affect how users engage with and interpret AI-generated health content, reinforcing the need for adaptive trust calibration mechanisms that account for these pre-existing cognitive processes.

In addition, this thesis extended the established distinctions between self-reported and physiological measures to the context of LLM-driven health communication. The work of this thesis demonstrated how discrepancies between self-reported trust and implicit responses (i.e., behavioral and physiological responses) manifest when users engage with LLM-generated health information. The difference underscored the complex cognitive processes involved in trust formation, where explicit judgments may not always align with the implicit reactions. Future AI-powered health systems can incorporate real-time trust assessment mechanisms that leverage such implicit signals, such as eye-tracking and cognitive load indicators, to detect trust fluctuations and dynamically adjust transparency, confidence displays, explanations, and interaction strategies. For example, if a user exhibits hesitation or skepticism, a trust-aware AI system could provide additional references, alternative perspectives, or reassurance features to support informed trust.

Furthermore, findings revealed that AI-labeled content triggers higher cognitive scrutiny, challenging the assumption that transparency always enhances trust. Instead, static transparency labels may introduce skepticism or cognitive dissonance, leading to inconsistent trust perceptions. These insights emphasize the need for context-sensitive transparency strategies that dynamically calibrate disclosures based on user expertise, engagement levels, and task criticality, ensuring users neither blindly trust nor unjustly dismiss LLM-generated health content, ultimately fostering more informed and balanced trust perceptions in LLM-driven health communication.

Ethical and Regulatory Considerations for Human-Centered AI in Digital Health Context

This thesis highlights the potential of LLMs in generating structured and explainable psychotherapeutic dialogues, but their application in sensitive health contexts raises ethical concerns. Unlike traditional rule-based AI used for task-oriented purposes, LLM-powered conversational AI engages directly with individuals in free, open-ended conversational interactions. Without proper regulatory safeguards, such as the EU AI Act [431] and GDPR [433], users may perceive harmfulness or develop over-reliance on AI-generated responses. To mitigate these risks, LLM-generated health advice must be explainable, traceable, and auditable to empower informed decision-making. This thesis proposed alignment approaches for AI-driven psychotherapy, yet broader regulations must ensure that explainability becomes a standard requirement in LLM-powered healthcare. Besides, accountability frameworks should define liability in cases where AI-generated content leads to misinterpretation, misinformation, or harm, especially in health contexts [372]. Establishing shared accountability between AI developers, healthcare providers, and regulatory bodies is essential to balance the benefits of AI-driven digital health solutions with responsible oversight. Additionally, fairness assessments must be integrated into regulatory frameworks, ensuring that AI is audited for biases in health content based on demographic and cultural factors [438].

Moreover, the integration of AI in healthcare introduces complex transparency challenges, particularly in balancing user awareness and trust calibration. While AI-powered health systems enhance accessibility and democratize medical knowledge, this thesis underscores the risks of over-reliance on AI or LLM-generated content and misguided trust in AI-driven health recommendations [439, 440]. Current regulatory frameworks often emphasize static AI disclosures, assuming that labeling online health information enhances transparency. While the work in this thesis does not implement adaptive transparency, we speculate that it may support informed trust by dynamically adjusting AI explanations and disclosures based on user expertise, engagement levels, and task criticality. However, such mechanisms also raise ethical and security concerns around manipulation, fairness, and user autonomy, underscoring the need for careful design and future validation. Future regulatory policies should define clear ethical boundaries for adaptive AI disclosures in healthcare, ensuring transparency about the non-human nature of interactions and preserving informed user consent. The research in this thesis also demonstrates the potential of behavioral and physiological sensing as implicit measures of trust in AI-driven health communication. While these techniques offer exploratory and valuable insights into user confidence and skepticism, they also introduce significant privacy and security concerns regarding AI's ability to monitor and interpret user behavior [59, 297, 355]. The ethical implications of AI-driven trust measurement extend beyond standard privacy issues, raising concerns about user autonomy, informed consent, and potential AI manipulation [365]. Regulatory frameworks should establish strict ethical boundaries to ensure transparent and responsible use of behavioral sensing methods and data. AI and LLM-powered systems should not exploit implicit trust markers to subtly manipulate user perceptions. Instead, consent-based mechanisms must be implemented, granting users control over data collection and AI adaptation features.

7

Beyond compliance and transparency, the role of AI in healthcare should be carefully positioned as an augmentative tool rather than a replacement for human expertise [441]. As AI systems become more sophisticated, health decision-making will increasingly involve hybrid human-AI collaboration, requiring policies that preserve human oversight rather than allowing full automation. This thesis emphasizes the importance of designing AI as assistants, providing recommendations while keeping final decision-making under human control. To support safe and effective AI integration into healthcare, regulatory guidelines should mandate shared decision-making models, where LLM-generated health information is explainable and subject to human validation. Additionally, clear guidelines on responsibility-sharing between AI developers, healthcare professionals, and policymakers are crucial to ensuring ethical and safe deployment of AI and LLMs in health contexts. Furthermore, human-centered AI policies should promote professional validation of AI-generated health content, requiring AI-driven systems to undergo rigorous evaluation in real-world settings before widespread deployment. Ethical standards should also prohibit fully autonomous AI decision-making in high-stakes health scenarios, reinforcing that AI should support rather than replace human judgment.

7.3 Limitations and Future Work

While this thesis advances the understanding of LLM-powered health intervention, trust in LLM-powered health information seeking, and human-AI collaboration in healthcare, several limitations should be acknowledged. These limitations inform directions for future research, which can build upon the findings to enhance the explainability and transparency of LLM-driven health applications or systems.

Limited Scope of LLM Alignment for Psychotherapeutic Health Intervention

One of the primary limitations of this research lies in the scope of LLM alignment for psychotherapy. While this thesis focused on aligning LLMs with expert-crafted dialogue scripts based on cognitive-behavioral therapy (CBT) [17] and motivational interviewing (MI) strategies [11, 64], MI and CBT are just two of many established counseling techniques. Other evidence-based therapies, such as dialectical behavior therapy (DBT) [436], follow distinct conversational structures that may not generalize well to the expertise-driven alignment (SSAG) developed here and vice versa. Even within MI, training data limitations pose challenges. The datasets (Chapter 3 and 4) used for training and evaluation, though valuable, may not fully capture the diversity of the therapeutical principles, linguistic variations, and real-world contextual factors. The annotated MI dataset (BiMISC, Chapter 3) for MISC coding was relatively small, potentially limiting the fine-grained prediction or classification of less common MI strategies such as complex reflections or nuanced affirmations. These gaps may lead to imbalanced model performance [442], highlighting the need for broader alignment approaches and more diverse, representative datasets to improve the robustness and adaptability of LLM-powered psychotherapy.

Another challenge is that LLM-generated psychotherapeutic dialogues remain static and session-specific, lacking approaches for longitudinal alignment across multiple therapy sessions. Real-world psychotherapy involves ongoing interactions where therapists dynamically adapt their behaviors based on the client's progress, emotional state, and evolving situation. Current LLM alignment methods do not account for the adaptive, multi-session nature of psychotherapy, limiting the applicability in sustained mental health intervention.

Future work should explore the multi-framework approaches for LLM alignment, integrating strategies or expertise from different therapeutic frameworks to enhance adaptability across diverse interventions. Expanding training datasets to include more varied therapeutic sessions and client demographics will also improve model robustness. Furthermore, incorporating memory-augmented LLMs or reinforcement learning from human feedback (RLHF) [175] could enable AI systems to track therapeutic progression over time, making LLM-powered psychotherapy more contextually aware and personalized.

Balancing Therapeutic Adherence and Empathy in LLM-Powered Psychotherapy

While aligning LLMs with MI strategies and expert-crafted dialogue scripts improves generative controllability and therapeutic explainability, a significant challenge remains in balancing adherence to therapeutic principles while maintaining the generation of emotionally nuanced, empathetic responses for more engaging conversations. Human therapists do not rely solely on predefined strategies, they dynamically adjust their language, tone, and emotional expressiveness based on subtle client cues, a level of adaptability that LLMs currently struggle to replicate.

In this thesis, Script-Strategy Aligned Generation (SSAG) followed MI strategies more effectively than non-aligned baselines, but this adherence sometimes came at the cost of emotional depth. For instance, LLM-generated reflections were often rated as appropriate and structured but lacked deeper empathetic resonance in MI sessions (Chapter 2), particularly when dealing with complex client statements involving ambivalence, distress, or personal struggles. Human therapists can intuitively adjust the depth of reflections, balancing technical accuracy with emotional engagement [11, 443], whereas LLMs may default to more surface-level paraphrasing, missing deeper therapeutic connection. Besides, LLMs lack true emotional understanding [437], relying on syntactic patterns and probabilistic estimations rather than genuine empathic reasoning like humans. Work from Syed et al. [22] further argues that AI perceives empathy differently than humans. Moreover, expertise-driven alignment assumes that predicting MI strategies before generating responses enhances explainability, yet therapeutic conversations are inherently fluid, requiring therapists to spontaneously adapt strategies based on client cues and emotional shifts. The rigid structure of LLMs may limit their ability to navigate unpredictable conversational turns, detect subtle emotional changes, or respond with warmth and affirmation, key elements of effective psychotherapy (Chapter 4). Additionally, while automatic metrics and human evaluations assess therapeutic adherence, they fail to capture deeper therapeutic impact, rapport-building, or the emotional resonance of dialogues. This highlights the need for more nuanced evaluation metrics, incorporating empathic accuracy, user engagement, and context-aware analysis to bridge the gap between structured alignment and natural, emotion-aware psychotherapeutic interactions.

Future research should explore aligning LLMs with emotion-aware signals, adaptive dialogue modeling, and (multimodal) sentiment analysis to enhance the naturalness and emotional intelligence of AI-driven psychotherapy. Developing dynamic, hybrid models that combine expert-guided therapeutic structures with more flexible, empathy-driven AI interactions will be critical for improving the practical application of LLMs in chatbot-delivered psychotherapy for health behavior intervention.

Challenges of Controlled Lab Studies in Capturing Real-World Trust Perceptions

Our investigation into understanding human trust perception in LLM-powered health information seeking also poses a few limitations (Chapter 5 and 6). While the lab-based experimental design allowed for in-depth exploration of trust formation in LLM-powered health information-seeking and reduced external confounders, it may not fully capture the complexity and spontaneity of real-world behaviors. In natural settings, individuals engage with health information dynamically, often revisiting sources, comparing multiple perspectives, and incorporating prior knowledge over time. The controlled lab environment may have influenced participants' responses, as the act of being observed can heighten cognitive scrutiny and skepticism toward AI-generated content, particularly due to societal caution regarding AI decision-making in sensitive domains. This observer effect from experimental environments could have led to more critical trust assessments than would occur in real-world, unmonitored interactions.

Additionally, in both survey and lab studies, trust was measured at a single point (Chapter 6), whereas in real-world interactions, trust evolves with repeated exposure. Initially, users may be skeptical of LLM-generated health information, but over time, trust could increase due to familiarity or decrease if inconsistencies or errors emerge. Repeated exposure may also lead to automation bias, where users begin to over-rely on LLM-generated responses without critical evaluation. Besides, explicit AI labeling may have different long-term effects, while AI-generated labels may initially trigger skepticism, sustained exposure could either normalize trust or reinforce concerns about credibility and reliability, especially if users associate AI-labeled content with misinformation risks or a lack of human expertise. Furthermore, we did not assess participants' actual decision-making behaviors after receiving health advice in the work of this thesis (Chapter 5 and 6). While trust in information was measured through self-reported ratings and physiological markers, it remains unclear whether these perceptions translate into real-world health decisions, such as adhering to AI-suggested treatments or seeking additional human consultation.

Future research should investigate whether trust levels differ for users who receive explicitly AI-labeled information versus unlabeled or hybrid AI-human content and whether labeling effects persist or fade over time. Beyond individual experiences with AI-generated health information, pre-existing trust in AI itself, shaped by media influence, cultural attitudes, and prior exposure to AI systems in other domains, may also significantly impact how users develop trust in LLM-powered health communication. Longitudinal studies are needed to examine how broader societal perceptions of AI influence trust in AI-generated health information, particularly as AI continues to evolve and integrate into high-stakes decision-making contexts like healthcare. Action-oriented trust can also be investigated using methods such as Ecological Momentary Assessment (EMA) [428], to track how users' decision-making evolves over time.

The Complexity of Interpreting Physiological and Behavioral Signals as Trust Indicators

The lab study (Chapter 6) employed physiological (i.e., ECG (heart rate variability)) and EDA (skin conductance)) and behavioral sensing methods (i.e., eye-tracking) to assess participants' implicit responses to received health information. While this method could provide valuable trust-related indicators, human trust is a complex and subjective phenomenon that cannot be fully captured through physiological spikes or gaze patterns alone. For instance, increased physiological activity could indicate trust-related cognitive engagement, but it could also reflect stress, information overload, or an unrelated emotional response. Without careful contextualization, such signals risk being misinterpreted, leading to overgeneralized or misleading conclusions about trust formation. Moreover, this thesis used self-reported trust ratings as a reference point to evaluate physiological and behavioral signals, yet such self-reports are themselves imperfect, subject to introspective bias. The observed divergence between self-reports and implicit responses raises questions about treating self-reports as definitive ground truths. Furthermore, physiological and behavioral responses were collected in a controlled lab environment, where external factors were minimized. However, trust can be influenced by situational factors in real-world settings, such as personal health concerns, situational pressure, and prior experiences with AI-based advice. Future studies should incorporate advanced trust measurement frameworks, combining physiological sensing, self-reported trust assessments, and behavioral indicators over extended periods to better capture the nuances of trust formation. Additionally, advancements in adaptive AI trust calibration mechanisms, where AI systems dynamically adjust transparency, explanations, and uncertainty disclosures based on real-time user engagement, could provide deeper insights into how users interact with LLM-generated health content over time.

Lastly, the use of physiological measurement in AI systems raises ethical concerns regarding data privacy, user consent, and potential AI overreach. Trust-aware AI systems must be designed with strong ethical safeguards, ensuring that behavioral and physiological data are used transparently and responsibly, with clear user control over data collection and interpretation. Future research should explore privacy-preserving methods and regulatory frameworks, such as those outlined in GDPR [433] and European AI Act [431], to establish ethical guidelines for using trust-aware AI in healthcare applications or systems.

7.4 Conclusion

This thesis explores two critical aspects of LLM-driven healthcare: (1) aligning large language models (LLMs) with domain expertise for health interventions, and (2) understanding trust in LLM-powered health communication for health information seeking and dissemination. Through computational modeling, empirical studies, and theoretical analysis, the research in this thesis advances the development of controllable, explainable, and transparent AI applications in healthcare. The findings provide valuable insights into how AI can be designed to support health interventions and health communication while addressing challenges related to alignment, transparency, and user trust perception.

The first research theme focused on aligning LLMs with domain expertise to improve their effectiveness in psychotherapy. While LLMs generate fluent and contextually appropriate dialogue, their lack of intrinsic adherence to therapeutic principles limits their therapeutical effectiveness. To address this, this research introduced an expertise-driven alignment approach (i.e., SSAG), where LLMs were guided to predict MI strategies before generating utterances. The results demonstrated significant improvements in therapeutic adherence, response coherence, and explainability, contributing to the development of structured and explainable LLM-powered psychotherapy applications.

The second research theme examined trust in LLM-powered health information seeking, a key factor influencing the adoption of health information. Findings revealed that interaction modality played a crucial role, with text-based interfaces evoking higher perceived trust than speech-based and embodied ones. Notably, participants reported greater trust in ChatGPT compared to traditional search engines (i.e., Google), reflecting the quality and coherence of LLM-generated responses in health contexts. Additionally, although users generally trusted LLM-generated content more than human-generated health information, human-labeled content was still perceived as more trustworthy, highlighting the influence of transparency on trust formation. Furthermore, behavioral and physiological data (e.g., eye-tracking and ECG signals) were used to capture implicit trust responses, offering insights into implicit reactions that may complement self-reported trust assessment. While self-reports served as an important reference point, the sensing data provided an additional layer of evidence to better understand participants' underlying trust perceptions. These findings suggested that future AI-powered health systems may explore adaptive trust calibration mechanisms that adjust explanations and transparency levels based on user engagement, though further research is needed to evaluate their effectiveness and ethical implications.

From a broader perspective, this thesis contributes to responsible AI integration in healthcare by addressing key challenges related to AI alignment, transparency, and trust calibration. The findings in this thesis offer both theoretical and practical implications: (1) demonstrating how expertise-driven alignment enhances LLM-powered health interventions, (2) providing insights into trust formation in LLM-powered health communication, and (3) informing regulatory and ethical considerations, advocating for context-sensitive transparency models, objective trust measurement, and AI-assisted decision-making that prioritizes human oversight. Ultimately, this thesis envisions a future where AI and human experts collaborate in the health context, with AI serving as a supportive, explainable, and transparent tool rather than an opaque decision-making system. By continuing to explore the intersection of AI alignment, transparency, and human trust, researchers and practitioners can contribute to the development of AI-driven applications that empower users, enhance therapeutic decision-making, and improve access to high-quality, personalized healthcare services.

Bibliography

[1] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of chatgpt: The history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.

[2] K. R. Chowdhary, *Natural Language Processing*. New Delhi: Springer India, 2020, pp. 603–649. [Online]. Available: https://doi.org/10.1007/978-81-322-3972-7_19

[3] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: a systematic review," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018.

[4] L. He, E. Başar, R. Wiers, M. Antheunis, and E. Krahmer, "Can chatbots help to motivate smoking cessation? a study on the effectiveness of motivational interviewing on engagement and therapeutic alliance," *BMC Public Health*, 04 2022.

[5] X. Sun, D. Casula, A. Navaratnam, A. Popp, F. Knopp, G. Busini, J. Wesołowski, M. V. Reeth, E. Reich, R. Wiers, and J. A. Bosch, "Virtual support for real-world movement: Using chatbots to overcome barriers to physical activity," in *HHAI 2023: Augmenting Human Intellect*, ser. Frontiers in Artificial Intelligence and Applications, vol. 368, 2023, pp. 201–214.

[6] F. Almusharraf, J. Rose, and P. Selby, "Engaging unmotivated smokers to move toward quitting: design of motivational interviewing–based chatbot through iterative interactions," *Journal of Medical Internet Research*, vol. 22, no. 11, p. e20251, 2020.

[7] L. He, E. Basar, R. W. Wiers, M. L. Antheunis, and E. Krahmer, "Can chatbots help to motivate smoking cessation? a study on the effectiveness of motivational interviewing on engagement and therapeutic alliance," *BMC Public Health*, vol. 22, no. 1, p. 726, Apr. 2022.

[8] W. Miller and S. Rollnick, "Motivational interviewing: Preparing people for change, 2nd ed." *Journal For Healthcare Quality*, vol. 25, p. 46, 05 2002.

[9] R. Martins and D. McNeil, "Review of motivational interviewing in promoting health behaviors," *Clinical psychology review*, vol. 29, pp. 283–93, 03 2009.

[10] *The Handbook of Behavior Change*, ser. Cambridge Handbooks in Psychology. Cambridge University Press, 2020.

[11] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc)," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2002.

[12] M. McTear, *Rule-Based Dialogue Systems: Architecture, Methods, and Tools*. Cham: Springer International Publishing, 2021, pp. 43–70. [Online]. Available: https://doi.org/10.1007/978-3-031-02176-3_2

[13] G. R. S. Silva and E. D. Canedo, "Towards user-centric guidelines for chatbot conversational design," *International Journal of Human–Computer Interaction*, vol. 40, no. 2, p. 98–120, Sep. 2022. [Online]. Available: http://dx.doi.org/10.1080/10447318.2022.2118244

[14] M. Urban and S. Mailey, "Conversation design: Principles, strategies, and practical application," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–3. [Online]. Available: https://doi.org/10.1145/3290607.3298821

[15] I. M. Rathi, S. Taylor, B. Bergen, and C. Jones, "Gpt-4 is judged more human than humans in displaced and inverted turing tests," in *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, and G. Mikros, Eds. Abu Dhabi, UAE: International Conference on Computational Linguistics, Jan. 2025, pp. 96–110. [Online]. Available: https://aclanthology.org/2025.genaidetect-1.7/

[16] J. E. Fischer, "Generative ai considered harmful," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3571884.3603756

[17] J. S. Beck, *Cognitive behavior therapy: Basics and beyond.* The Guilford Press, 2021.

[18] W. Miller and S. Rollnick, "Motivational interviewing: Preparing people for change, 2nd ed." *Journal For Healthcare Quality*, vol. 25, p. 46, 05 2002.

[19] D. Alperstein and L. Sharpe, "The efficacy of motivational interviewing in adults with chronic pain: A meta-analysis and systematic review," *The Journal of Pain*, vol. 17, no. 4, pp. 393–403, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1526590015009578

[20] J. S. Beck, *Cognitive behavior therapy: Basics and beyond.* Guilford Press, 2011.

[21] Z. Iftikhar, S. Ransom, A. Xiao, and J. Huang, "Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt," 2024. [Online]. Available: https://arxiv.org/abs/2409.02244

[22] S. Syed, Z. Iftikhar, A. W. Xiao, and J. Huang, "Machine and human understanding of empathy in online peer support: A cognitive behavioral approach," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642034

[23] R. J. W. Cline and K. M. Haynes, "Consumer health information seeking on the Internet: the state of the art," *Health Education Research*, vol. 16, no. 6, pp. 671–692, 12 2001. [Online]. Available: https://doi.org/10.1093/her/16.6.671

[24] E. Sillence, P. Briggs, P. R. Harris, and L. Fishwick, "How do patients evaluate and make use of online health information?" *Social Science and Medicine*, vol. 64, no. 9, pp. 1853–1862, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0277953607000160

[25] "Products - data briefs - number 482 - oxctober 2023 — cdc.gov," https://www.cdc.gov/nchs/products/databriefs/db482.htm, 2022.

[26] Eurostat, "survey on the use of ict in households and by individuals," 2022. [Online]. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220406-1

[27] A. El Ali, K. P. Venkatraj, S. Morosoli, L. Naudts, N. Helberger, and P. Cesar, "Transparent ai disclosure obligations: Who, what, when, where, why, how," in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613905.3650750

[28] A. N. Desai, D. Ruidera, J. M. Steinbrink, B. Granwehr, and D. H. Lee, "Misinformation and disinformation: The potential disadvantages of social media in infectious disease and how to combat them," *Clinical Infectious Diseases*, vol. 74, pp. e34–e39, 2022.

[29] J. Dalton, S. Fischer, P. Owoicho, F. Radlinski, F. Rossetto, J. R. Trippas, and H. Zamani, "Conversational information seeking: Theory and application," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3455–3458. [Online]. Available: https://doi.org/10.1145/3477495.3532678

[30] X. Sun, Y. Liu, J. De Wit, J. A. Bosch, and Z. Li, "Trust by interface: How different user interfaces shape human trust in health information from large language models," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613905.3650837

[31] Y. Jung, C. Chen, E. Jang, and S. S. Sundar, "Do we trust chatgpt as much as google search and wikipedia?" in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613905.3650862

[32] R. Xu, Y. Feng, and H. Chen, "Chatgpt vs. google: A comparative study of search performance and user experience," 2023. [Online]. Available: https://arxiv.org/abs/2307.01135

[33] X. Sun, R. Ma, X. Zhao, Z. Li, J. Lindqvist, A. E. Ali, and J. A. Bosch, "Trusting the search: Unraveling human trust in health information from google and chatgpt," 2024. [Online]. Available: https://arxiv.org/abs/2403.09987

[34] "Kff-health misinformation tracking poll: Artificial intelligence and health information | kff — kff.org," https://www.kff.org/health-information-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/, 2025.

[35] H. Choung, P. David, and A. Ross, "Trust in ai and its role in the acceptance of ai technologies," *International Journal of Human–Computer Interaction*, vol. 39, no. 9, pp. 1727–1739, 2023. [Online]. Available: https://doi.org/10.1080/10447318.2022.2050543

[36] A. Gupta, D. Basu, R. Ghantasala, S. Qiu, and U. Gadiraju, "To trust or not to trust: How a conversational interface affects trust in a decision support system," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3531–3540. [Online]. Available: https://doi.org/10.1145/3485447.3512248

[37] C. Wang, S. Liu, H. Yang, J. Guo, Y. Wu, and J. Liu, "Ethical considerations of using chatgpt in health care," *J Med Internet Res*, vol. 25, p. e48009, Aug 2023. [Online]. Available: https://www.jmir.org/2023/1/e48009

[38] J. Marecos, D. Tude Graça, F. Goiana-da Silva, H. Ashrafian, and A. Darzi, "Source credibility labels and other nudging interventions in the context of online health misinformation: A systematic literature review," *Journalism and Media*, vol. 5, no. 2, pp. 702–717, 2024. [Online]. Available: https://www.mdpi.com/2673-5172/5/2/46

[39] A. Kaplan, T. Kessler, and P. Hancock, "How trust is defined and its use in human-human and human-machine interaction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 64, pp. 1150–1154, 12 2020.

[40] J. B. Rotter, "A new scale for the measurement of interpersonal trust1," *J. Pers.*, vol. 35, no. 4, pp. 651–665, Dec. 1967.

[41] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995. [Online]. Available: http://www.jstor.org/stable/258792

[42] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A systematic literature review of user trust in ai-enabled systems: An hci perspective," *International Journal of Human–Computer Interaction*, vol. 40, no. 5, pp. 1251–1266, 2024. [Online]. Available: https://doi.org/10.1080/10447318.2022.2138826

[43] J. D. Lee and K. A. See, "Trust in automation: designing for appropriate reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[44] O. Vereschak, G. Bailly, and B. Caramiaux, "How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, Oct. 2021. [Online]. Available: https://doi.org/10.1145/3476068

[45] S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi, "Trust in ai: progress, challenges, and future directions," *Humanities and Social Sciences Communications*, vol. 11, no. 1, p. 1568, Nov 2024. [Online]. Available: https://doi.org/10.1057/s41599-024-04044-8

[46] T. Lucassen and J. M. Schraagen, "Trust in wikipedia: How users trust information from an unknown source," in *Proceedings of the 4th Workshop on Information Credibility*, ser. WICOW '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 19–26. [Online]. Available: https://doi.org/10.1145/1772938.1772944

[47] E. Sillence, P. Briggs, L. Fishwick, and P. Harris, "Guidelines for developing trust in health websites," in *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, ser. WWW '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 1026–1027. [Online]. Available: https://doi.org/10.1145/1062745.1062851

[48] B. R. Bates, S. Romina, R. Ahmed, and D. Hopson, "The effect of source credibility on consumers' perceptions of the quality of health information on the internet," *Medical informatics and the Internet in medicine*, vol. 31, no. 1, pp. 45–52, 2006.

[49] X. Wang, Y. Yang, D. Tao, and T. Zhang, "The impact of AI transparency and reliability on Human-AI collaborative Decision-Making," in *Artificial Intelligence, Social Computing and Wearable Technologies. AHFE (2023) International Conference. AHFE Open Access*, W. Karwowski and T. Ahram, Eds. USA: AHFE International, 2023, vol. 113.

[50] R. F. Kizilcec, "How much information? effects of transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 2390–2395. [Online]. Available: https://doi.org/10.1145/2858036.2858402

[51] S. Eckhardt, N. Kühl, M. Dolata, and G. Schwabe, "A survey of ai reliance," 2024. [Online]. Available: https://arxiv.org/abs/2408.03948

[52] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. Bernstein, and R. Krishna, "Explanations can reduce overreliance on ai systems during decision-making," 2023. [Online]. Available: https://arxiv.org/abs/2212.06823

[53] J. Chen, S. Mishler, and B. Hu, "Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 5, pp. 463–473, 2021.

[54] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw, "Measurement of trust in automation: A narrative review and reference guide," *Front Psychol*, vol. 12, p. 604977, Oct. 2021.

[55] F. Anvari, E. Efendic, J. Olsen, R. Arslan, M. Elson, and I. Schneider, "Bias in self-reports: An initial elevation phenomenon," *Social Psychological and Personality Science*, vol. 14, no. 6, pp. 727–737, Aug. 2023.

[56] I. B. Ajenaghughrure, S. D. C. Sousa, and D. Lamas, "Measuring trust with psychophysiological signals: A systematic mapping study of approaches used," *Multimodal Technologies and Interaction*, vol. 4, no. 3, 2020. [Online]. Available: https://www.mdpi.com/2414-4088/4/3/63

[57] K. Akash, W.-L. Hu, N. Jain, and T. Reid, "A classification model for sensing human trust in machines using eeg and gsr," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 4, nov 2018. [Online]. Available: https://doi.org/10.1145/3132743

[58] J. Z. Lim, J. Mountstephens, and J. Teo, "Eye-tracking feature extraction for biometric machine learning," *Front Neurorobot*, vol. 15, p. 796895, Feb. 2022.

[59] F. Chiossi, E. R. Stepanova, B. Tag, M. Perusquia-Hernandez, A. Kitson, A. Dey, S. Mayer, and A. El Ali, "Physiochi: Towards best practices for integrating physiological signals in hci," in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613905.3636286

[60] R. Tiwari, R. Kumar, S. Malik, T. Raj, and P. Kumar, "Analysis of heart rate variability and implication of different factors on heart rate variability," *Curr Cardiol Rev*, vol. 17, no. 5, p. e160721189770, 2021.

[61] M. Ahmad and A. Alzahrani, "Crucial clues: Investigating psychophysiological behaviors for measuring trust in human-robot interaction," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 135–143. [Online]. Available: https://doi.org/10.1145/3577190.3614148

[62] E. Babaei, B. Tag, T. Dingler, and E. Velloso, "A critique of electrodermal activity practices at chi," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445370

[63] C. P. D. R. S. Jannet M. de Jonge, Gerard M. Schippers, "The motivational interviewing skill code: Reliability and a critical appraisal," *Cambridge University Press*, 01 2005.

[64] R. S. Shah, F. Holt, S. A. Hayati, A. Agarwal, Y.-C. Wang, R. E. Kraut, and D. Yang, "Modeling motivational interviewing strategies on an online peer-to-peer counseling platform," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, nov 2022. [Online]. Available: https://doi.org/10.1145/3555640

[65] OpenAI. [Online]. Available: https://openai.com/chatgpt/

[66] K. Holmqvist, M. Nystrom, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. United States: Oxford University Press, 2011.

[67] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford press, 2012.

[68] R. K. Martins and D. W. McNeil, "Review of motivational interviewing in promoting health behaviors," *Clinical Psychology Review*, vol. 29, no. 4, pp. 283–293, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0272735809000099

[69] S. Park, J. Choi, S. Lee, C. Oh, C. Kim, S. La, J. Lee, and B. Suh, "Designing a chatbot for a brief motivational interview on stress management: Qualitative case study," *Journal of Medical Internet Research*, vol. 21, no. 4, p. e12231, Apr 2019. [Online]. Available: https://www.jmir.org/2019/4/e12231/

[70] B. Xu and Z. Zhuang, "Survey on psychotherapy chatbots," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 7, p. e6170, 2022.

[71] J. Zhang, Y. J. Oh, P. Lange, Z. Yu, and Y. Fukuoka, "Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: Viewpoint," *Journal of Medical Internet Research*, vol. 22, no. 9, p. e22845, Sep 2020. [Online]. Available: https://www.jmir.org/2020/9/e22845

[72] L. He, E. Başar, R. W. Wiers, M. L. Antheunis, and E. Krahmer, "Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance," *BMC Public Health*, vol. 22, no. 1, p. 726, 2022.

[73] D. Min, V. Perez-Rosas, K. Resnicow, and R. Mihalcea, "Verve: Template-based reflective rewriting for motivational interviewing," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023, pp. 10 289–10 302.

[74] L. He, E. Başar, E. Krahmer, R. Wiers, and M. Antheunis, "Effectiveness and user experience of a smoking cessation chatbot: A mixed-methods study comparing motivational interviewing and confrontational counseling," *Journal of Medical Internet Research*, vol. 26, p. e53134, 2024.

[75] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, no. 1, p. 65–170, jan 2018. [Online]. Available: https://dl.acm.org/doi/10.5555/3241691.3241693

[76] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, ser. Studies in Natural Language Processing. Cambridge University Press, 2000.

[77] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, "A survey of natural language generation," *ACM Computing Surveys*, vol. 55, no. 8, Dec. 2022. [Online]. Available: https://doi.org/10.1145/3554727

[78] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2023.

[79] B. Peng, C. Zhu, C. Li, X. Li, J. Li, M. Zeng, and J. Gao, "Few-shot natural language generation for task-oriented dialog," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 172–182. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.17

[80] A. Ferrario and N. Biller-Andorno, "Large language models in medical ethics: useful but not expert," *Journal of Medical Ethics*, vol. 50, no. 9, pp. 653–654, 2024. [Online]. Available: https://jme.bmj.com/content/50/9/653

[81] H. Li, J. Moon, S. Purkayastha, L. Celi, H. Trivedi, and J. Gichoya, "Ethics of large language models in medicine and medical research," *The Lancet Digital Health*, vol. 5, 04 2023.

[82] F. Bianchi and J. Zou, "Large language models are vulnerable to bait-and-switch attacks for generating harmful content," 2024.

[83] OpenAI, "Gpt-4 technical report," 2023.

[84] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[85] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," 2022.

[86] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. Reforgiato Recupero, and D. Riboni, "Anno-MI: A dataset of expert-annotated counselling dialogues," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6177–6181.

[87] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, D. Reforgiato Recupero, and D. Riboni, "Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues," *Future Internet*, vol. 15, no. 3, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/3/110

[88] J. Passmore, "Motivational interviewing techniques reflective listening," in *Coaching Practiced*. John Wiley & Sons Ltd., 2022, pp. 251–255.

[89] K. Resnicow and F. McMaster, "Motivational interviewing: moving from why to how with autonomy support," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 9, pp. 1–9, 2012.

[90] J. Dieter, T. Wang, A. T. Chaganty, G. Angeli, and A. X. Chang, "Mimic and rephrase: Reflective listening in open-ended dialogue," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, M. Bansal and A. Villavicencio, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 393–403. [Online]. Available: https://aclanthology.org/K19-1037

[91] S. Shen, C. Welch, R. Mihalcea, and V. Pérez-Rosas, "Counseling-style reflection generation using generative pretrained transformers with augmented context," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, Jul. 2020, pp. 10–20. [Online]. Available: https://aclanthology.org/2020.sigdial-1.2

[92] I. Ahmed, E. Keilty, C. Cooper, P. Selby, and J. Rose, "Generation and classification of motivational-interviewing-style reflections for smoking behaviour change using few-shot learning with transformers." 2022.

[93] A. Brown, A. T. Kumar, O. Melamed, I. Ahmed, Y. H. Wang, A. Deza, M. Morcos, L. Zhu, M. Maslej, N. Minian, V. Sujaya, J. Wolff, O. Doggett, M. Iantorno, M. Ratto, P. Selby, and J. Rose, "A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study," *JMIR Mental Health*, vol. 10, p. e49132, Oct 2023. [Online]. Available: https://doi.org/10.2196/49132

[94] A. Brown, J. Zhu, M. Abdelwahab, A. Dong, C. Wang, and J. Rose, “Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model,” 2024.

[95] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, “Human evaluation of automatically generated text: Current trends and best practice guidelines,” *Computer Speech & Language*, vol. 67, p. 101151, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082030084X

[96] A. Celikyilmaz, E. Clark, and J. Gao, “Evaluation of text generation: A survey,” 2021.

[97] A. B. Sai, A. K. Mohankumar, and M. M. Khapra, “A survey of evaluation metrics used for nlg systems,” *ACM Computing Surveys*, vol. 55, no. 2, Jan. 2022. [Online]. Available: https://doi.org/10.1145/3485766

[98] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[99] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[100] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[101] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[102] J. Novikova, O. Dušek, and V. Rieser, “RankME: Reliable human ratings for natural language generation,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 72–78. [Online]. Available: https://aclanthology.org/N18-2012/

[103] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, “Manual for the motivational interviewing skill code (MISC),” *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[104] J. M. de Jonge, G. M. Schippers, and C. P. Schaap, “The motivational interviewing skill code: Reliability and a critical appraisal,” *Behavioural and Cognitive Psychotherapy*, vol. 33, no. 3, p. 285–298, 2005.

[105] A. Liesenfeld, A. Lopez, and M. Dingemanse, “Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators,” in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI ’23. Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3571884.3604316

[106] R. K. Maurya, “Using ai based chatbot chatgpt for practicing counseling skills through role-play,” *Journal of Creativity in Mental Health*, vol. 19, no. 4, pp. 513–528, 2024.

[107] M. Shanahan, K. McDonell, and L. Reynolds, “Role play with large language models,” *Nature*, vol. 623, pp. 493–498, 2023.

[108] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” in *8th International Conference on Learning Representations, ICLR*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[109] T. Sellam, D. Das, and A. Parikh, “BLEURT: Learning robust metrics for text generation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892. [Online]. Available: https://aclanthology.org/2020.acl-main.704/

[110] Z. Wu, S. Balloccu, E. Reiter, R. Helaoui, D. Reforgiato Recupero, and D. Riboni, “Are experts needed? on human evaluation of counselling reflection generation,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6906–6930. [Online]. Available: https://aclanthology.org/2023.acl-long.382/

[111] E. Başar, I. Hendrickx, E. Krahmer, G. de Bruijn, and T. Bosse, “To what extent are large language models capable of generating substantial reflections for motivational interviewing counseling chatbots? A human evaluation,” in *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*. Association for Computational Linguistics, 2024, pp. 41–52.

[112] T. Boardman, D. Catley, J. E. Grobe, T. D. Little, and J. S. Ahluwalia, “Using motivational interviewing with smokers: Do therapist behaviors relate to engagement and therapeutic alliance?” *Journal of Substance Abuse Treatment*, vol. 31, no. 4, pp. 329–339, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0740547206001462

[113] A. See, S. Roller, D. Kiela, and J. Weston, “What makes a good conversation? how controllable attributes affect human judgments,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1702–1723. [Online]. Available: https://aclanthology.org/N19-1170/

[114] J. Amidei, P. Piwek, and A. Willis, “The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations,” in *Proceedings of the 12th International Conference on Natural Language Generation*, K. van Deemter, C. Lin, and H. Takamura, Eds. Tokyo, Japan: Association for Computational Linguistics, Oct.–Nov. 2019, pp. 397–402. [Online]. Available: https://aclanthology.org/W19-8648/

[115] R. Herbrich, T. Minka, and T. Graepel, “Trueskill™: A bayesian skill rating system,” in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2006/file/f44ee263952e65b3610b8ba51229d1f9-Paper.pdf

[116] E. Başar, D. Balaji, L. He, I. Hendrickx, E. Krahmer, G. de Bruijn, and T. Bosse, “HyLECA: A framework for developing hybrid long-term engaging controlled conversational agents,” in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI ’23. Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3571884.3604404

[117] W. Miller and S. Rollnick, “Motivational interviewing: Preparing people for change (2nd ed.),” *Journal For Healthcare Quality*, vol. 25, p. 46, 05 2002.

[118] R. Martins and D. McNeil, “Review of motivational interviewing in promoting health behaviors,” *Clinical psychology review*, vol. 29, pp. 283–93, 03 2009.

[119] L. Tavabi, T. Tran, K. Stefanov, B. Borsari, J. Woolley, S. Scherer, and M. Soleymani, “Analysis of behavior classification in motivational interviewing,” vol. 2021, 06 2021, pp. 110–115.

[120] C. P. D. R. S. Jannet M. de Jonge, Gerard M. Schippers, “The motivational interviewing skill code: Reliability and a critical appraisal,” *Cambridge University Press*, 01 2005.

[121] D. Atkins, M. Steyvers, Z. Imel, and P. Smyth, “Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification,” *Implementation science : IS*, vol. 9, p. 49, 04 2014.

[122] J. Cao, M. Tanana, Z. Imel, E. Poitras, D. Atkins, and V. Srikumar, “Observing dialogue in therapy: Categorizing and forecasting behavioral codes,” 01 2019, pp. 5599–5611.

[123] M. Tanana, K. Hallgren, Z. Imel, D. Atkins, and V. Srikumar, “A comparison of natural language processing methods for automated coding of motivational interviewing,” *Journal of Substance Abuse Treatment*, vol. 65, 01 2016.

[124] F. Klonek, V. Quera, and S. Kauffeld, “Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers?” *Computers in Human Behavior*, vol. 44, 03 2015.

[125] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, L. An, K. J. Goggin, and D. Catley, “Predicting counselor behaviors in motivational interviewing encounters,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1128–1137. [Online]. Available: https://aclanthology.org/E17-1106

[126] D. Can, D. Atkins, and S. Narayanan, “A dialog act tagging approach to behavioral coding: a case study of addiction counseling conversations,” 09 2015, pp. 339–343.

[127] B. Xiao, D. Can, J. Gibson, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, “Behavioral coding of therapist language in addiction counseling using recurrent neural networks,” 09 2016, pp. 908–912.

[128] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, D. Reforgiato Recupero, and D. Riboni, “Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues,” *Future Internet*, vol. 15, no. 3, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/3/110

[129] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, and L. An, "Building a motivational interviewing dataset," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 42–51. [Online]. Available: https://aclanthology.org/W16-0305

[130] W. Zixiu, B. Simone, K. Vivek, H. Rim, R. Ehud, R. R. Diego, and R. Daniele, "Anno-mi: A dataset of expert-annotated counselling dialogues," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6177–6181.

[131] A. Welivita and P. Pu, "Curating a large-scale motivational interviewing dataset using peer support forums," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3315–3330. [Online]. Available: https://aclanthology.org/2022.coling-1.293

[132] W. Deng, J. Pei, Z. Ren, Z. Chen, and P. Ren, "Intent-calibrated self-training for answer selection in open-domain dialogues," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1232–1249, 2023.

[133] G. Yan, J. Pei, P. Ren, Z. Ren, and M. de Rijke, "Remedi: Resources for multi-domain, multi-service, medical dialogues," 2022, international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22).

[134] W. Deng, J. Pei, K. Kong, Z. Chen, F. Wei, Y. Li, Z. Ren, Z. Chen, and P. Ren, "Syllogistic reasoning for legal judgment analysis," in *Empirical Methods in Natural Language Processing (EMNLP'23)*, 2023.

[135] R. Chew, J. Bollenbacher, M. Wenger, J. Speer, and A. Kim, "Llm-assisted content analysis: Using large language models to support deductive coding," 06 2023.

[136] S. D. Paoli, "Can large language models emulate an inductive thematic analysis of semi-structured interviews? an exploration and provocation on the limits of the approach and the model," 2023.

[137] R. H. Tai, L. R. Bentley, X. Xia, J. M. Sitt, S. C. Fankhauser, A. M. Chicas-Mosier, and B. M. Monteith, "Use of large language models to aid analysis of textual data," *bioRxiv*, 2023. [Online]. Available: https://www.biorxiv.org/content/early/2023/07/19/2023.07.17.549361

[138] T. B. Moyers, L. N. Rowell, J. K. Manuel, D. Ernst, and J. M. Houck, "The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity," *J Subst Abuse Treat*, vol. 65, pp. 36–42, Jan. 2016.

[139] A. Street, "Alexander street," 2023. [Online]. Available: https://alexanderstreet.com/

[140] D. Korngiebel and S. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery," *npj Digital Medicine*, vol. 4, 12 2021.

[141] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, and P.-Y. Oudeyer, "Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, ser. IUI '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, pp. 75–78. [Online]. Available: https://doi.org/10.1145/3581754.3584136

[142] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[143] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3816–3830. [Online]. Available: https://aclanthology.org/2021.acl-long.295

[144] Google, "Google translate," 2023. [Online]. Available: https://translate.google.com

[145] J. Opitz and S. Burst, "Macro f1 and macro f1," 2021.

[146] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022.

[147] X. Sun, D. Casula, A. Navaratnam, A. Popp, F. Knopp, G. Busini, J. Wesołowski, M. V. Reeth, E. Reich, R. Wiers, and J. A. Bosch, "Virtual support for real-world movement: Using chatbots to overcome barriers to physical activity," in *HHAI 2023: Augmenting Human Intellect*, ser. Frontiers in Artificial Intelligence and Applications, vol. 368, 2023, pp. 201–214.

[148] S. Park, J. Choi, S. Lee, C. Oh, C. Kim, S. La, J. Lee, and B. Suh, "Designing a chatbot for a brief motivational interview on stress management: Qualitative case study," *J Med Internet Res*, vol. 21, no. 4, p. e12231, Apr 2019. [Online]. Available: https://doi.org/10.2196/12231

[149] B. Xu and Z. Zhuang, "Survey on psychotherapy chatbots," *Concurrency and Computation: Practice and Experience*, 12 2020.

[150] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, "A survey of natural language generation," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, Dec. 2022. [Online]. Available: http://dx.doi.org/10.1145/3554727

[151] E. Basar, D. Balaji, L. He, I. Hendrickx, E. Krahmer, G.-J. de Bruijn, and T. Bosse, "Hyleca: A framework for developing hybrid long-term engaging controlled conversational agents," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3571884.3604404

[152] X. Sun, E. Krahmer, J. De Wit, R. Wiers, and J. A. Bosch, "Plug and play conversations: The micro-conversation scheme for modular development of hybrid conversational agent," in *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, p. 50–55. [Online]. Available: https://doi.org/10.1145/3584931.3606998

[153] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[154] H. Na, Y. Hua, Z. Wang, T. Shen, B. Yu, L. Wang, W. Wang, J. Torous, and L. Chen, "A survey of large language models in psychotherapy: Current landscape and future directions," 2025. [Online]. Available: https://arxiv.org/abs/2502.11095

[155] M. V. Heinz, D. M. Mackin, B. M. Trudeau, S. Bhattacharya, Y. Wang, H. A. Banta, A. D. Jewett, A. J. Salzhauer, T. Z. Griffin, and N. C. Jacobson, "Randomized trial of a generative ai chatbot for mental health treatment," *NEJM AI*, vol. 2, no. 4, p. AIoa2400802, 2025. [Online]. Available: https://ai.nejm.org/doi/full/10.1056/AIoa2400802

[156] A. Ferguson, L.-A. Dinh-Williams, and Z. Segal, "Mindfulness-based cognitive therapy," in *Handbook of cognitive behavioral therapy: Overview and approaches (Vol. 1).* Washington: American Psychological Association, 2021, pp. 595–615.

[157] ""should" statements," https://www.psychologytools.com/resource/should-statements, Feb. 2023, accessed: 2024-10-28.

[158] W. Yang, A. Fang, R. S. Shah, Y. Mathur, D. Yang, H. Zhu, and R. E. Kraut, "What makes digital support effective? how therapeutic skills affect clinical well-being," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW1, apr 2024. [Online]. Available: https://doi.org/10.1145/3641029

[159] R. S. Shah, F. Holt, S. A. Hayati, A. Agarwal, Y.-C. Wang, R. E. Kraut, and D. Yang, "Modeling motivational interviewing strategies on an online peer-to-peer counseling platform," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, nov 2022. [Online]. Available: https://doi.org/10.1145/3555640

[160] X. Sun, X. Tang, A. El Ali, Z. Li, P. Ren, J. de Wit, J. Pei, and J. A.Bosch, "Rethinking the alignment of psychotherapy dialogue generation with motivational interviewing strategies," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 1983–2002. [Online]. Available: https://aclanthology.org/2025.coling-main.136/

[161] S. C. Hayes, K. D. Strosahl, and K. G. Wilson, *Acceptance and commitment therapy: The process and practice of mindful change*. The Guilford Press, 2012.

[162] S. J. Diller and J. Passmore, "Defining digital coaching: a qualitative inductive approach," *Frontiers in Psychology*, vol. 14, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263716095

[163] A. Sell, Christer, M. Helmefalk, and L. Marcusson, "Digital coaching to support university students' physical activity," 2019.

[164] C. Maher, C. Davis, R. Curtis, C. Short, and K. Murphy, "A physical activity and diet program delivered by artificially-intelligent virtual health coach: a case-controlled proof-of-concept study (preprint)," *JMIR mHealth and uHealth*, vol. 8, 12 2019.

[165] J. de Wit and A. Braggaar, "Tilbot: A visual design platform to facilitate open science research into conversational user interfaces," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3571884.3604403

[166] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3816–3830. [Online]. Available: https://aclanthology.org/2021.acl-long.295

[167] E. Basar, I. Hendrickx, E. Krahmer, G.-J. Bruijn, and T. Bosse, "To what extent are large language models capable of generating substantial reflections for motivational interviewing counseling chatbots? a human evaluation," in *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, N. Soni, L. Flek, A. Sharma, D. Yang, S. Hooker, and H. A. Schwartz, Eds. TBD: ACL, Aug. 2024, pp. 41–52. [Online]. Available: https://aclanthology.org/2024.hucllm-1.4

[168] C. Pelau, D.-C. Dabija, and I. Ene, "What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry," *Computers in Human Behavior*, vol. 122, p. 106855, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563221001783

[169] D. Demszky, D. Yang, D. S. Yeager, C. J. Bryan, M. Clapper, S. Chandhok, J. C. Eichstaedt, C. Hecht, J. Jamieson, M. Johnson, M. Jones, D. Krettek-Cobb, L. Lai, N. JonesMitchell, D. C. Ong, C. S. Dweck, J. J. Gross, and J. W. Pennebaker, "Using large language models in psychology," *Nature Reviews Psychology*, vol. 2, no. 11, pp. 688–701, Nov 2023. [Online]. Available: https://doi.org/10.1038/s44159-023-00241-5

[170] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445188

[171] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1137–1155, Mar. 2003.

[172] A. Welivita and P. Pu, "Boosting distress support dialogue responses with motivational interviewing strategy," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023.

[173] M. R. Kabir, R. M. Sultan, I. H. Asif, J. I. Ahad, F. Rahman, M. R. Amin, N. Mohammed, and S. Rahman, "Beyond labels: Aligning large language models with human-like reasoning," 2024. [Online]. Available: https://arxiv.org/abs/2408.11879

[174] natolambert Nathan Lambert; lcastricato Louis Castricato; lvwerra Leandro von Werra; Dahoas1 Alex Havrilla, "Illustrating reinforcement learning from human feedback (rlhf)," 12 2022. [Online]. Available: https://huggingface.co/blog/rlhf

[175] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 04 2022.

[176] Y. Deng, W. Zhang, Y. Yuan, and W. Lam, "Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations," 2023.

[177] A. S. Rao, A. Khandelwal, K. Tanmay, U. Agarwal, and M. Choudhury, "Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 370–13 388. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.892/

[178] E. Tennant, S. Hailes, and M. Musolesi, “Moral alignment for llm agents,” 2025. [Online]. Available: https://arxiv.org/abs/2410.01639

[179] P. Gupta, C. Jiao, Y.-T. Yeh, S. Mehri, M. Eskenazi, and J. P. Bigham, “Instructdial: Improving zero and few-shot generalization in dialogue through instruction tuning,” 2022. [Online]. Available: https://arxiv.org/abs/2205.12673

[180] H. Zhang, C. Tang, T. Loakman, B. Yang, S. Goetze, and C. Lin, “Cadge: Context-aware dialogue generation enhanced with graph-structured knowledge aggregation,” 2024. [Online]. Available: https://arxiv.org/abs/2305.06294

[181] J. M. Kwak, M. Kim, and S. J. Hwang, “Context-dependent instruction tuning for dialogue response generation,” 2023. [Online]. Available: https://arxiv.org/abs/2311.07006

[182] Q. Tu, Y. Li, J. Cui, B. Wang, J.-R. Wen, and R. Yan, “Misc: A mixed strategy-aware model integrating comet for emotional support conversation,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 308–319.

[183] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, “Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support,” *Nature Machine Intelligence*, vol. 5, no. 1, pp. 46–57, Jan 2023. [Online]. Available: https://doi.org/10.1038/s42256-022-00593-2

[184] X. Sun, X. Tang, A. E. Ali, Z. Li, X. Shen, P. Ren, J. de Wit, J. Pei, and J. A. Bosch, “Chain-of-strategy planning with llms: Aligning the generation of psychotherapy dialogue with strategy in motivational interviewing,” 2024. [Online]. Available: https://arxiv.org/abs/2408.06527

[185] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, Aug 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06291-2

[186] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, “Fine-tuning language models from human preferences,” 2020. [Online]. Available: https://arxiv.org/abs/1909.08593

[187] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, pp. 1 – 35, 2021.

[188] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” 2023.

[189] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic chain of thought prompting in large language models,” 2022.

[190] H. Wang, R. Wang, F. Mi, Y. Deng, Z. Wang, B. Liang, R. Xu, and K.-F. Wong, “Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 047–12 064. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.806

[191] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, “Tree of thoughts: deliberate problem solving with large language models,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2024.

[192] E. Basar, X. Sun, I. Hendrickx, J. de Wit, T. Bosse, G.-J. De Bruijn, J. A. Bosch, and E. Krahmer, “How well can large language models reflect? a human evaluation of LLM-generated reflections for motivational interviewing dialogues,” in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 1964–1982. [Online]. Available: https://aclanthology.org/2025.coling-main.135/

[193] A. Kermani, V. Perez-Rosas, and V. Metsis, “A systematic evaluation of llm strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. rag,” 2025. [Online]. Available: https://arxiv.org/abs/2503.24307

[194] M. E. Fonteyn, B. Kuipers, and S. J. Grobe, “A description of think aloud method and protocol analysis,” *Qualitative Health Research*, vol. 3, no. 4, pp. 430–441, 1993. [Online]. Available: https://doi.org/10.1177/104973239300300403

[195] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, “Rasa: Open source language understanding and dialogue management,” 2017. [Online]. Available: https://arxiv.org/abs/1712.05181

[196] O. et al., “Gpt-4 technical report,” 2024.

[197] OpenAI, “Fine-tuning,” 2024, [Accessed 17-09-2024]. [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning

[198] D. C. Kozen, *Depth-First and Breadth-First Search*. New York, NY: Springer New York, 1992, pp. 19–24. [Online]. Available: https://doi.org/10.1007/978-1-4612-4400-4_4

[199] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2006.14799

[200] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," 2018. [Online]. Available: https://arxiv.org/abs/1703.09902

[201] A. Braggaar, C. Liebrecht, E. van Miltenburg, and E. Krahmer, "Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations," 2024. [Online]. Available: https://arxiv.org/abs/2312.13871

[202] R. Rubin and M. Martin, "Development of a measure of interpersonal competence," *Communication Research Reports*, vol. 11, pp. 33–44, 06 1994.

[203] H. L. O' Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form," *International Journal of Human-Computer Studies*, vol. 112, pp. 28–39, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581918300041

[204] M. B. Madson, R. S. Mohn, J. A. Schumacher, and A. S. Landry, "Measuring client experiences of motivational interviewing during a lifestyle intervention," *Meas Eval Couns Dev*, vol. 48, no. 2, pp. 140–151, Apr. 2015.

[205] P. Henson, P. Peck, and J. Torous, "Considering the therapeutic alliance in digital mental health interventions," *Harv. Rev. Psychiatry*, vol. 27, no. 4, pp. 268–273, 2019.

[206] S. Borsci, A. Malizia, M. Schmettow, F. van der Velde, G. Tariverdiyeva, D. Balaji, and A. Chamberlain, "The chatbot usability scale: the design and pilot of a usability scale for interaction with ai-based conversational agents," *Personal and Ubiquitous Computing*, vol. 26, pp. 1–25, 02 2022.

[207] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175–191, May 2007.

[208] S. S. SHAPIRO and M. B. WILK, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, dec 1965. [Online]. Available: https://doi.org/10.1093/biomet/52.3-4.591

[209] H. Arsham and M. Lovric, *Bartlett's Test*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 87–88. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_132

[210] J. W. Hardin and J. M. Hilbe, *Generalized estimating equations, second edition*, 2nd ed. Philadelphia, PA: Chapman & Hall/CRC, Dec. 2012.

[211] D. Rey and M. Neuhäuser, *Wilcoxon-Signed-Rank Test*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1658–1659. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_616

[212] W. Haynes, *Bonferroni Correction*. New York, NY: Springer New York, 2013, pp. 154–154. [Online]. Available: https://doi.org/10.1007/978-1-4419-9863-7_1213

[213] J. H. Zar, "Spearman rank correlation," *Encyclopedia of Biostatistics*, vol. 7, 2005.

[214] S. Elo and H. Kyngäs, "The qualitative content analysis process," *Journal of advanced nursing*, vol. 62, no. 1, pp. 107–115, 2008.

[215] "ATLAS.Ti," https://atlasti.com, Sep. 2024, accessed: 2024-9-11.

[216] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024.

[217] H. Matsumoto, A. Taniguchi, and J. Nishida, "A revised self-determined motivation scale for exercise with integrated regulation inclusion," *Journal of Health Psychology Research*, vol. 34, 05 2021.

[218] vicuna, "lmsys/vicuna-13b-v1.5 · Hugging Face — huggingface.co," https://huggingface.co/lmsys/vicuna-13b-v1.5.

[219] J. Cao, M. Tanana, Z. Imel, E. Poitras, D. Atkins, and V. Srikumar, "Observing dialogue in therapy: Categorizing and forecasting behavioral codes," 01 2019, pp. 5599–5611.

[220] W. Zixiu, B. Simone, K. Vivek, H. Rim, R. Ehud, R. R. Diego, and R. Daniele, "Anno-mi: A dataset of expert-annotated counselling dialogues," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6177–6181.

[221] X. Sun, J. Pei, J. d. Wit, M. Aliannejadi, E. Krahmer, J. T. Dobber, and J. A. Bosch, "Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 5609–5621. [Online]. Available: https://aclanthology.org/2024.lrec-main.498

[222] Wikipedia, "Entropy (information theory)," May 2024. [Online]. Available: https://en.wikipedia.org/wiki/Entropy_(information_theory)

[223] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, M. R. Costa-jussà and E. Alfonseca, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 37–42. [Online]. Available: https://aclanthology.org/P19-3007

[224] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381. [Online]. Available: https://aclanthology.org/P19-1534

[225] Y. K. Lee, J. Suh, H. Zhan, J. J. Li, and D. C. Ong, "Large language models produce responses perceived to be empathic," 2024. [Online]. Available: https://arxiv.org/abs/2403.18148

[226] M. Luo, C. J. Warren, L. Cheng, H. M. Abdul-Muhsin, and I. Banerjee, "Assessing empathy in large language models with real-world physician-patient interactions," 2024. [Online]. Available: https://arxiv.org/abs/2405.16402

[227] M. Roshanaei, R. Rezapour, and M. S. El-Nasr, "Talk, listen, connect: Navigating empathy in human-ai interactions," 2024. [Online]. Available: https://arxiv.org/abs/2409.15550

[228] A. Cuadra, M. Wang, L. A. Stein, M. F. Jung, N. Dell, D. Estrin, and J. A. Landay, "The illusion of empathy? notes on displays of emotion in human-computer interaction," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642336

[229] M. Zhang, Z. Cheng, S. T. R. Shiu, J. Liang, C. Fang, Z. Ma, L. Fang, and S. J. Wang, "Towards human-centred ai-co-creation: A three-level framework for effective collaboration between human and ai," in *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, p. 312–316. [Online]. Available: https://doi.org/10.1145/3584931.3607008

[230] E. Mosqueira Rey, E. Hernandez Pereira, D. Alonso Rios, J. Bobes Bascaran, and A. Fernandez Leal, "Human-in-the-loop machine learning: a state of the art," 08 2022.

[231] X. Sun, J. A. Bosch, J. De Wit, and E. Krahmer, "Human-in-the-loop interaction for continuously improving generative model in conversational agent for behavioral intervention," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, ser. IUI '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, p. 99–101. [Online]. Available: https://doi.org/10.1145/3581754.3584142

[232] L. Eltahawy, T. Essig, N. Myszkowski, and L. Trub, "Can robots do therapy?: Examining the efficacy of a cbt bot in comparison with other behavioral intervention technologies in alleviating mental health symptoms," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, p. 100035, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S294988212300035X

[233] S. Karkosz, R. Szymański, K. Sanna, and J. Michałowski, "Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: Randomized controlled trial," *JMIR Form Res*, vol. 8, p. e47960, Mar 2024. [Online]. Available: https://formative.jmir.org/2024/1/e47960

[234] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR Ment Health*, vol. 4, no. 2, p. e19, Jun 2017. [Online]. Available: http://mental.jmir.org/2017/2/e19/

[235] H. Kibirige and L. Ghemri, "Trust ranking of medical websites," in *Proceedings of the 4th ACM Conference on Data and Application Security and Privacy*, ser. CODASPY '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 151–154. [Online]. Available: https://doi.org/10.1145/2557547.2557584

[236] MAYO CLINIC. (2023) Mayo clinic. [Online]. Available: https://www.mayoclinic.org/

[237] P. N. Venkit, P. Laban, Y. Zhou, Y. Mao, and C.-S. Wu, "Search engines in an ai era: The false promise of factual and verifiable source-cited responses," 2024. [Online]. Available: https://arxiv.org/abs/2410.22349

[238] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[239] R. K. Garg, V. L. Urs, A. A. Agrawal, S. K. Chaudhary, V. Paliwal, and S. K. Kar, "Exploring the role of chatgpt in patient care (diagnosis and treatment) and medical research: A systematic review," *medRxiv*, 2023. [Online]. Available: https://www.medrxiv.org/content/early/2023/06/14/2023.06.13.23291311

[240] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine," *New England Journal of Medicine*, vol. 388, no. 13, pp. 1233–1239, 2023, pMID: 36988602. [Online]. Available: https://doi.org/10.1056/NEJMsr2214184

[241] S. S. Biswas, "Role of chatgpt in public health," *Annals of Biomedical Engineering*, vol. 51, p. 868–869, 2023. [Online]. Available: https://doi.org/10.1007/s10439-023-03172-7

[242] S. Lin, L. Lin, C. Hou, B. Chen, J. Li, and S. Ni, "Empathy-based communication framework for chatbots: A mental health chatbot application and evaluation," in *Proceedings of the 11th International Conference on Human-Agent Interaction*, ser. HAI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 264–272. [Online]. Available: https://doi.org/10.1145/3623809.3623865

[243] L. Vega, E. Montague, and T. DeHart, "Trust in health websites: A review of an emerging field," in *Proceedings of the 1st ACM International Health Informatics Symposium*, ser. IHI '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 700–709. [Online]. Available: https://doi.org/10.1145/1882992.1883100

[244] E. Sillence, P. Briggs, L. Fishwick, and P. Harris, "Trust and mistrust of online health sites," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 663–670. [Online]. Available: https://doi.org/10.1145/985692.985776

[245] T. Chin, "Patients put trust in internet health information," *American Medical News*, vol. 45, no. 23, pp. 26–27, 2002.

[246] V. Zieglmeier and A. M. Lehene, "Designing trustworthy user interfaces," in *Proceedings of the 33rd Australian Conference on Human-Computer Interaction*, ser. OzCHI '21. New York, NY, USA: Association for Computing Machinery, 2022, p. 182–189. [Online]. Available: https://doi.org/10.1145/3520495.3520525

[247] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 07 2018. [Online]. Available: https://doi.org/10.1093/jamia/ocy072

[248] M. Rheu, J. Y. Shin, W. Peng, and J. Huh-Yoo, "Systematic review: Trust-building factors and implications for conversational agent design," *International Journal of Human-Computer Interaction*, vol. 37, pp. 1–16, 09 2020.

[249] U. Valley, "Uncanny valley - wikipedia — en.wikipedia.org," https://en.wikipedia.org/wiki/Uncanny_valley, 2024.

[250] K. E. Arleen Salles and M. Farisco, "Anthropomorphism in ai," *AJOB Neuroscience*, vol. 11, no. 2, pp. 88–95, 2020, pMID: 32228388. [Online]. Available: https://doi.org/10.1080/21507740.2020.1740350

[251] A. Placani, "Anthropomorphism in ai: hype and fallacy," *AI and Ethics*, Feb 2024. [Online]. Available: https://doi.org/10.1007/s43681-024-00419-4

[252] D. McKnight, M. Carter, J. Thatcher, and P. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on Management Information Systems*, vol. 2, pp. 12–32, 06 2011.

[253] ——, "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on Management Information Systems*, vol. 2, pp. 12–32, 06 2011.

[254] F. C. Johnson, J. E. Rowley, and L. Sbaffi, "Modelling trust formation in health information contexts," *Journal of Information Science*, vol. 41, pp. 415 – 429, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206454953

[255] Y. Zhang, N. Suhaimi, N. Yongsatianchot, J. D. Gaggiano, M. Kim, S. A. Patel, Y. Sun, S. Marsella, J. Griffin, and A. G. Parker, "Shifting trust: Examining how trust and distrust emerge, transform, and collapse in covid-19 information seeking," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491102.3501889

[256] J. Liu, Y. Zhang, and Y. Kim, "Consumer health information quality, credibility, and trust: An analysis of definitions, measures, and conceptual dimensions," in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 197–210. [Online]. Available: https://doi.org/10.1145/3576840.3578331

[257] H. Singal and S. Kohli, "Intellectualizing trust for medical websites," in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, ser. ICTCS '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2905055.2905293

[258] M. Dutta-Bergman *et al.*, "Trusted online sources of health information: differences in demographics, health beliefs, and health-information orientation," *Journal of medical Internet research*, vol. 5, no. 3, p. e893, 2003.

[259] A. Yuan, A. Coenen, E. Reif, and D. Ippolito, "Wordcraft: Story writing with large language models," *27th International Conference on Intelligent User Interfaces*, 2022.

[260] F. Davis and F. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, pp. 319–, 09 1989.

[261] E. Sillence, J. M. Blythe, P. Briggs, and M. Moss, "A revised model of trust in internet-based health information and advice: Cross-sectional questionnaire study." *Journal of medical Internet research*, 2019.

[262] Y. Guo, "Digital trust and the reconstruction of trust in the digital society: An integrated model based on trust theory and expectation confirmation theory," *Digit. Gov.: Res. Pract.*, vol. 3, no. 4, dec 2022. [Online]. Available: https://doi.org/10.1145/3543860

[263] L. Daraz, A. S. Morrow, O. J. Ponce, B. Beuschel, M. H. Farah, A. Katabi, M. Alsawas, A. M. Majzoub, R. Benkhadra, M. O. Seisa, J. F. Ding, L. Prokop, and M. H. Murad, "Can patients trust online health information? a meta-narrative systematic review addressing the quality of health information on the internet," *J Gen Intern Med*, vol. 34, no. 9, pp. 1884–1891, Jun. 2019.

[264] M. Metzger and A. Flanagin, "Credibility and trust of information in online environments: The use of cognitive heuristics," *Journal of Pragmatics*, vol. 59, p. 210–220, 12 2013.

[265] E. Sillence, P. Briggs, P. Harris, and L. Fishwick, "A framework for understanding trust factors in web-based health advice," *Int. J. Hum.-Comput. Stud.*, vol. 64, no. 8, p. 697–713, aug 2006. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2006.02.007

[266] B. Friedman, J. C. Thomas, J. Grudin, C. Nass, H. Nissenbaum, M. Schlager, and B. Shneiderman, "Trust me, i'm accountable: Trust and accountability online," in *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 79–80. [Online]. Available: https://doi.org/10.1145/632716.632766

[267] G. Bansal and M. Warkentin, "Do you still trust? the role of age, gender, and privacy concern on trust after insider data breaches," *SIGMIS Database*, vol. 52, no. 4, p. 9–44, dec 2022. [Online]. Available: https://doi.org/10.1145/3508484.3508487

[268] C. di Sciascio, E. Veas, J. Barria-Pineda, and C. Culley, "Understanding the effects of control and transparency in searching as learning," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ser. IUI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 498–509. [Online]. Available: https://doi.org/10.1145/3377325.3377524

[269] E. Ul Haque, M. M. H. Khan, and M. A. A. Fahim, "The nuanced nature of trust and privacy control adoption in the context of google," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3581387

[270] M. Seckler, S. Heinz, S. Forde, A. N. Tuch, and K. Opwis, "Trust and distrust on the web," *Comput. Hum. Behav.*, vol. 45, no. C, p. 39–50, apr 2015. [Online]. Available: https://doi.org/10.1016/j.chb.2014.11.064

[271] M. De Choudhury, M. R. Morris, and R. W. White, "Seeking and sharing health information online: comparing search engines and social media," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014, pp. 1365–1376.

[272] Y.-M. Kim, "Is seeking health information online different from seeking general information online?" *J. Inf. Sci.*, vol. 41, no. 2, p. 228–241, apr 2015. [Online]. Available: https://doi.org/10.1177/0165551514561669

[273] National Institutes of Health. (2023) National institutes of health. [Online]. Available: https://www.nih.gov

[274] C. Corritore, S. Wiedenbeck, B. Kracher, and R. Marble, “Online trust and health information websites,” *International Journal of Technology and Human Interaction*, vol. 8, 01 2007.

[275] Q. Gao and G.-J. Houben, “A framework for trust establishment and assessment on the web of data,” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1097–1098. [Online]. Available: https://doi.org/10.1145/1772690.1772822

[276] B. W. Hesse, D. E. Nelson, G. L. Kreps, R. T. Croyle, N. K. Arora, B. K. Rimer, and K. Viswanath, “Trust and sources of health information: the impact of the internet and its implications for health care providers: findings from the first health information national trends survey,” *Archives of internal medicine*, vol. 165, no. 22, pp. 2618–2624, 2005.

[277] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, pp. 1 – 35, 2021.

[278] T. Rietz and A. Maedche, “Cody: An ai-based system to semi-automate coding for qualitative research,” 05 2021.

[279] Y. Gil and D. Artz, “Towards content trust of web resources,” in *Proceedings of the 15th International Conference on World Wide Web*, ser. WWW ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 565–574. [Online]. Available: https://doi.org/10.1145/1135777.1135861

[280] H. Häußler, “Do users trust search engines? and if so, why? developing a trust measure and applying it in an experiment,” in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, ser. CHIIR ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 471–474. [Online]. Available: https://doi.org/10.1145/3576840.3578280

[281] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” vol. 21, no. 1, jan 2020.

[282] A. V. Raju Vaishya, Anoop Misra, “Chatgpt: Is this version good for healthcare and research?” *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 17, pp. 1871–4021, 2023. [Online]. Available: https://doi.org/10.1016/j.dsx.2023.102744

[283] V. Hristidis, N. Ruggiano, E. L. Brown, S. R. R. Ganta, and S. Stewart, “Chatgpt vs google for queries related to dementia and other cognitive decline: Comparison of results,” *J Med Internet Res*, vol. 25, p. e48966, Jul 2023. [Online]. Available: https://www.jmir.org/2023/1/e48966

[284] L. Van Bulck and P. Moons, “What if your patient switches from dr. google to dr. chatgpt? a vignette-based survey of the trustworthiness, value, and danger of chatgpt-generated responses to health questions,” *European Journal of Cardiovascular Nursing*, p. zvad038, 04 2023. [Online]. Available: https://doi.org/10.1093/eurjcn/zvad038

[285] A. Følstad, C. B. Nordheim, and C. A. Bjørkli, “What makes users trust a chatbot for customer service? an exploratory interview study,” in *Internet Science*, S. S. Bodrunova, Ed. Cham: Springer International Publishing, 2018, pp. 194–208.

[286] M. Jang and T. Lukasiewicz, “Consistency analysis of chatgpt,” 2023.

[287] R. H. Wortham and A. Theodorou, “Robot transparency, trust and utility,” *Connection Science*, vol. 29, no. 3, pp. 242–248, 2017. [Online]. Available: https://doi.org/10.1080/09540091.2017.1313816

[288] A. L. Fruhling and S. M. Lee, “The influence of user interface usability on rural consumers’ trust of e-health services,” *International journal of electronic healthcare*, vol. 2, no. 4, pp. 305–321, 2006.

[289] K. Torning and H. Oinas-Kukkonen, “Persuasive system design : State of the art and future directions,” vol. 350, 04 2009, p. 30.

[290] E. Luger and A. Sellen, “”like having a really bad pa”: The gulf between user expectation and experience of conversational agents,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 5286–5297. [Online]. Available: https://doi.org/10.1145/2858036.2858288

[291] L. Zhang, J. Yu, S. Zhang, L. Li, Y. Zhong, G. Liang, Y. Yan, Q. Ma, F. Weng, F. Pan, J. Li, R. Xu, and Z. Lan, “Unveiling the impact of multi-modal interactions on user engagement: A comprehensive evaluation in ai-driven conversations,” 2024. [Online]. Available: https://arxiv.org/abs/2406.15000

[292] I. Syed, M. Baart, and J. Vroomen, “The multimodal trust effects of face, voice, and sentence content,” *Multisens Res*, vol. 37, no. 2, pp. 125–141, Apr. 2024.

[293] C. F. Lam, C. Lee, and Y. Sui, “Say it as it is: Consequences of voice directness, voice politeness, and voicer credibility on voice endorsement,” *J. Appl. Psychol.*, vol. 104, no. 5, pp. 642–658, May 2019.

[294] F. Gaiser and S. Utz, "Is hearing really believing? the importance of modality for perceived message credibility during information search with smart speakers," *Journal of Media Psychology*, 06 2023.

[295] M. L. Lupetti, E. Hagens, W. Van Der Maden, R. Steegers-Theunissen, and M. Rousian, "Trustworthy embodied conversational agents for healthcare: A design exploration of embodied conversational agents for the periconception period at erasmus mc," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3571884.3597128

[296] S. Oviatt, *Multimodal Interaction, Interfaces, and Analytics*. Cham: Springer International Publishing, 2020, pp. 1–29. [Online]. Available: https://doi.org/10.1007/978-3-319-27648-9_22-1

[297] G. Friedland and M. C. Tschantz, *Privacy Concerns of Multimodal Sensor Systems*. Association for Computing Machinery and Morgan & Claypool, 2019, p. 659–704. [Online]. Available: https://doi.org/10.1145/3233795.3233813

[298] S. Yadav, D. Gupta, and D. Demner-Fushman, "Chq-summ: A dataset for consumer healthcare question summarization," 2022.

[299] Yahoo. (2023) L6 - yahoo! answers comprehensive questions and answers version 1.0. [Online]. Available: https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11

[300] S. Jessup, T. Schneider, G. Alarcon, T. Ryan, and A. Capiola, *The Measurement of the Propensity to Trust Automation*, 06 2019, pp. 476–489.

[301] J. E. Rowley, F. C. Johnson, and L. Sbaffi, "Students' trust judgements in online health information seeking," *Health Informatics Journal*, vol. 21, pp. 316 – 327, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:21888204

[302] E. R. Girden, *ANOVA: Repeated measures.*, ser. Sage University papers. Quantitative applications in the social sciences, Vol. 84. Thousand Oaks, CA, US: Sage Publications, Inc, 1992.

[303] A. Ross and V. L. Willson, *Paired Samples T-Test*. Rotterdam: SensePublishers, 2017, pp. 17–19. [Online]. Available: https://doi.org/10.1007/978-94-6351-086-8_4

[304] V. Braun and V. Clarke, *Thematic analysis.* American Psychological Association, 2012.

[305] N. McDonald, S. Schoenebeck, and A. Forte, "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–23, 2019.

[306] Wikipedia, "Filter bubble - wikipedia," September 2023, (Accessed on 09/29/2023). [Online]. Available: https://en.wikipedia.org/wiki/Filter_bubble#:~:text=A%20filter%20bubble%20or%20ideological,can%20result%20from%20personalized%20searches.

[307] A. Carolus, M. J. Koch, S. Straka, M. E. Latoschik, and C. Wienrich, "Mails - meta ai literacy scale: Development and testing of an ai literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies," *Computers in Human Behavior: Artificial Humans*, vol. 1, no. 2, p. 100014, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949882123000142

[308] C. Teixeira Lopes and E. Ramos, "Studying how health literacy influences attention during online information seeking," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 283–291. [Online]. Available: https://doi.org/10.1145/3343413.3377966

[309] J. R. Lewis, B. S. Utesch, and D. E. Maher, "Umux-lite: When there's no time for the sus," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 2099–2102. [Online]. Available: https://doi.org/10.1145/2470654.2481287

[310] V. Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions," *Decision Sciences - DECISION SCI*, vol. 39, pp. 273–315, 05 2008.

[311] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[312] M. J. Lindstrom and D. M. Bates, "Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1014–1022, 1988. [Online]. Available: https://doi.org/10.1080/01621459.1988.10478693

[313] D. A. Robb, J. Lopes, M. I. Ahmad, P. E. McKenna, X. Liu, K. S. Lohan, and H. F. Hastie, “Seeing eye to eye: trustworthy embodiment for task-based conversational agents,” *Frontiers in Robotics and AI*, vol. 10, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261447242

[314] Z. He, “From eliza to chatgpt: The evolution of chatbots in public health,” *XRDS*, vol. 29, no. 3, p. 59, apr 2023. [Online]. Available: https://doi.org/10.1145/3589658

[315] A. Haleem, M. Javaid, and R. P. Singh, “An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges,” *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 2, no. 4, p. 100089, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772485923000066

[316] F. M. Hafizoglu and S. Sen, “Reputation based trust in human-agent teamwork without explicit coordination,” in *Proceedings of the 6th International Conference on Human-Agent Interaction*, ser. HAI ’18. New York, NY, USA: Association for Computing Machinery, 2018, p. 238–245. [Online]. Available: https://doi.org/10.1145/3284432.3284454

[317] OpenAI, “Chatgpt search,” https://openai.com/index/introducing-chatgpt-search/, 2024.

[318] R. Xu, Y. Feng, and H. Chen, “Chatgpt vs. google: A comparative study of search performance and user experience,” 2023.

[319] C. N. Harrington and L. Egede, “Trust, comfort and relatability: Understanding black older adults’ perceptions of chatbot design for health information seeking,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3580719

[320] M. Luria, G. Hoffman, and O. Zuckerman, “Comparing social robot, screen and voice interfaces for smart-home control,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 580–628. [Online]. Available: https://doi.org/10.1145/3025453.3025786

[321] J. Schaffer, J. O’Donovan, J. Michaelis, A. Raglin, and T. Höllerer, “I can do better than your ai: Expertise and explanations,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 240–251. [Online]. Available: https://doi.org/10.1145/3301275.3302308

[322] P. S. Raju and M. D. Reilly, “Product familiarity and information processing strategies: An exploratory investigation,” *Journal of Business Research*, vol. 8, no. 2, pp. 187–212, 1980.

[323] E. Kuang, E. Jahangirzadeh Soure, M. Fan, J. Zhao, and K. Shinohara, “Collaboration with conversational ai assistants for ux evaluation: Questions and how to ask them (voice vs. text),” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–15.

[324] D. Ellis, D. Cox, and K. Hall, “A comparison of the information seeking patterns of researchers in the physical and social sciences,” *Journal of documentation*, vol. 49, no. 4, pp. 356–369, 1993.

[325] M. Bokhari and F. Hasan, “Multimodal information retrieval: Challenges and future trends,” *International Journal of Computer Applications (0975 – 8887)*, vol. 74, pp. 9–12, 07 2013.

[326] E. Bruno, J. Kludas, and S. Marchand-Maillet, “Combining multimodal preferences for multimedia information retrieval,” in *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, ser. MIR ’07. New York, NY, USA: Association for Computing Machinery, 2007, p. 71–78. [Online]. Available: https://doi.org/10.1145/1290082.1290095

[327] Y. Deldjoo, J. R. Trippas, and H. Zamani, “Towards multi-modal conversational information seeking,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1577–1587. [Online]. Available: https://doi.org/10.1145/3404835.3462806

[328] E. Rusnandi, E. Winarko, and S. Azhari, “A survey on multimodal information retrieval approach,” in *2020 International Conference on Smart Technology and Applications (ICoSTA)*, 2020, pp. 1–6.

[329] S. Rafael, “Multimodality, naturalness and transparency in affective computing for hci,” in *Design, User Experience, and Usability. Interaction Design*, A. Marcus and E. Rosenzweig, Eds. Cham: Springer International Publishing, 2020, pp. 521–531.

[330] S. S. Sundar, “The main model : A heuristic approach to understanding technology effects on credibility,” 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:17588424

[331] J. Kim, K. Merrill Jr., K. Xu, and S. Kelly, "Perceived credibility of an ai instructor in online education: The role of social presence and voice features," *Computers in Human Behavior*, vol. 136, p. 107383, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563222002059

[332] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples, "Voice interfaces in everyday life," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–12. [Online]. Available: https://doi.org/10.1145/3173574.3174214

[333] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan, "Embodiment in conversational interfaces: Rea," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 520–527. [Online]. Available: https://doi.org/10.1145/302979.303150

[334] A. Antle, P. Marshall, and E. van den Hoven, "Workshop on embodied interaction: Theory and practice in hci," 05 2011, pp. 5–8.

[335] J. Cassell, T. Bickmore, H. Vilhjálmsson, and H. Yan, "More than just a pretty face: Affordances of embodiment," in *Proceedings of the 5th International Conference on Intelligent User Interfaces*, ser. IUI '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 52–59. [Online]. Available: https://doi.org/10.1145/325737.325781

[336] S. Olafsson, D. Parmar, E. Kimani, T. K. O'Leary, and T. Bickmore, " 'more like a person than reading text in a machine' : Characterizing user choice of embodied agents vs. conventional guis on smartphones," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411763.3451664

[337] J. Cassell, "Embodied conversational agents: Representation and intelligence in user interfaces," *AI Magazine*, vol. 22, no. 4, p. 67, Dec. 2001. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1593

[338] J. Kim and I. Im, "Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents," *Computers in Human Behavior*, vol. 139, p. 107512, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563222003326

[339] M. Li and A. Suh, "Machinelike or humanlike? a literature review of anthropomorphism in ai-enabled technology," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, ser. Proceedings of the Annual Hawaii International Conference on System Sciences, Jan. 2021, pp. 4053–4062, research Unit(s) information for this publication is provided by the author(s) concerned.; 54th Hawaii International Conference on System Sciences (HICSS 2021), HICSS-54 ; Conference date: 04-01-2021 Through 08-01-2021. [Online]. Available: https://scholarspace.manoa.hawaii.edu/handle/10125/72112

[340] J. Cassell, "Embodied conversational agents: Representation and intelligence in user interfaces," *AI Magazine*, vol. 22, no. 4, p. 67, Dec. 2001. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1593

[341] W. Street, "Llm theory of mind and alignment: Opportunities and risks," 2024. [Online]. Available: https://arxiv.org/abs/2405.08154

[342] C. Pelau, L.-E. Anica-Popa, I. Bojescu, and M. Niculescu, "Are men more affected by ai anthropomorphism? comparative research on the perception of ai human-like characteristics between genders," 05 2022.

[343] P. Carlbring, H. Hadjistavropoulos, A. Kleiboer, and G. Andersson, "A new era in internet interventions: The advent of chat-gpt and ai-assisted therapist guidance," *Internet Interventions*, vol. 32, 2023.

[344] J. E. H. Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, "Human versus artificial intelligence," *Frontiers in Artificial Intelligence*, vol. 4, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2021.622364

[345] M. Shanahan, K. McDonell, and L. Reynolds, "Role-play with large language models," 2023.

[346] R. K. Maurya, "Using ai based chatbot chatgpt for practicing counseling skills through role-play," Sep 2023. [Online]. Available: psyarxiv.com/s47jb

[347] C. Esterwood and L. P. Robert, "Personality in healthcare human robot interaction (h-hri): A literature review and brief critique," in *Proceedings of the 8th International Conference on Human-Agent Interaction*, ser. HAI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 87–95. [Online]. Available: https://doi.org/10.1145/3406499.3415075

[348] W. Cai, Y. Jin, and L. Chen, "Impacts of personal characteristics on user trust in conversational recommender systems," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491102.3517471

[349] Y. Li, K. Li, H. Ning, X. Xia, Y. Guo, C. Wei, J. Cui, and B. Wang, "Towards an online empathetic chatbot with emotion causes," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2041–2045. [Online]. Available: https://doi.org/10.1145/3404835.3463042

[350] A. R. Rahmanti, H.-C. Yang, B. S. Bintoro, A. A. Nursetyo, M. S. Muhtar, S. Syed-Abdul, and Y.-C. J. Li, "Slimme, a chatbot with artificial empathy for personal weight management: System design and finding," *Frontiers in Nutrition*, vol. 9, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnut.2022.870775

[351] B. Liu and S. S. Sundar, "Should machines express sympathy and empathy? experiments with a health advice chatbot," *Cyberpsychology, Behavior, and Social Networking*, vol. 21, no. 10, pp. 625–636, 2018, pMID: 30334655. [Online]. Available: https://doi.org/10.1089/cyber.2018.0110

[352] J. Festerling and I. Siraj, "Anthropomorphizing technology: A conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants," *Integrative Psychological and Behavioral Science*, vol. 56, no. 3, pp. 709–738, Sep 2022. [Online]. Available: https://doi.org/10.1007/s12124-021-09668-y

[353] C. S. Indrit Troshani, Sally Rao Hill and D. Arthur, "Do we trust in ai? role of anthropomorphism and intelligence," *Journal of Computer Information Systems*, vol. 61, no. 5, pp. 481–491, 2021. [Online]. Available: https://doi.org/10.1080/08874417.2020.1788473

[354] F. M. Calisto, J. a. Fernandes, M. Morais, C. Santiago, J. a. M. Abrantes, N. Nunes, and J. C. Nascimento, "Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3580682

[355] C. Katsini, Y. Abdrabou, G. E. Raptis, M. Khamis, and F. Alt, "The role of eye gaze in security and privacy applications: Survey and future hci research directions," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–21. [Online]. Available: https://doi.org/10.1145/3313831.3376840

[356] M. Carey, L. Crucianelli, C. Preston, and A. Fotopoulou, "The effect of visual capture towards subjective embodiment within the full body illusion," *Scientific Reports*, vol. 9, no. 1, p. 2889, Feb. 2019.

[357] L. L. Lott, F. B. Spengler, T. Stächele, B. Schiller, and M. Heinrichs, "EmBody/EmFace as a new open tool to assess emotion recognition from body and face expressions," *Scientific Reports*, vol. 12, no. 1, p. 14165, Aug. 2022.

[358] J. Wei, B. Tag, J. R. Trippas, T. Dingler, and V. Kostakos, "What could possibly go wrong when interacting with proactive smart speakers? a case study using an esm application," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491102.3517432

[359] Q. Liao and S. S. Sundar, "Designing for responsible trust in ai systems: A communication perspective," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1257–1268. [Online]. Available: https://doi.org/10.1145/3531146.3533182

[360] new Bing. (2023) new bing. [Online]. Available: https://www.bing.com/new?setlang=en&sid=06C46094C75F62702B3A7336C628637F

[361] Wikipwdia, "Hallucination (artificial intelligence) - wikipedia," https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence), September 2023, (Accessed on 10/10/2023).

[362] F. Walker, M. Favetta, L. Hasker, and R. Walker, "They prefer humans! experimental measurement of student trust in chatgpt," in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613905.3650955

[363] I. Rae, "The effects of perceived ai use on content perceptions," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642076

[364] D. Johnson, R. Goodman, J. Patrinely, C. Stone, E. Zimmerman, R. Donald, S. Chang, S. Berkowitz, A. Finn, E. Jahangir, E. Scoville, T. Reese, D. Friedman, J. Bastarache, Y. van der Heijden, J. Wright, N. Carter, M. Alexander, J. Choe, and L. Wheless, "Assessing the accuracy and reliability of ai-generated medical responses: An evaluation of the chat-gpt model," *Research square*, 02 2023.

[365] M. Reis, F. Reis, and W. Kunde, "Influence of believed AI involvement on the perception of digital medical advice," *Nature Medicine*, Jul. 2024.

[366] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organ. Behav. Hum. Decis. Process.*, vol. 151, pp. 90–103, Mar. 2019.

[367] Y. Yin, N. Jia, and C. J. Wakslak, "Ai can help people feel heard, but an ai label diminishes this impact," *Proceedings of the National Academy of Sciences*, vol. 121, no. 14, p. e2319112121, 2024. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2319112121

[368] S. Shekar, P. Pataranutaporn, C. Sarabu, G. A. Cecchi, and P. Maes, "People over trust ai-generated medical responses and view them to be as valid as doctors, despite low accuracy," 2024. [Online]. Available: https://arxiv.org/abs/2408.15266

[369] Q. V. Liao, H. Subramonyam, J. Wang, and J. Wortman Vaughan, "Designerly understanding: Information needs for model transparency to support design ideation for ai-powered user experience," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3580652

[370] J. Wang, "Gaze behavior, skin conductance, and trust in automation," August 2018. [Online]. Available: http://essay.utwente.nl/76357/

[371] O. Vereschak, F. Alizadeh, G. Bailly, and B. Caramiaux, "Trust in ai-assisted decision making: Perspectives from those behind the system and those for whom the decision is made," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642018

[372] B. Friedman, J. C. Thomas, J. Grudin, C. Nass, H. Nissenbaum, M. Schlager, and B. Shneiderman, "Trust me, i'm accountable: Trust and accountability online," in *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 79–80. [Online]. Available: https://doi.org/10.1145/632716.632766

[373] WebMD. [Online]. Available: https://www.webmd.com/

[374] T. Bickmore, A. Gruber, and R. Picard, "Establishing the computer–patient working alliance in automated health behavior change interventions," *Patient Education and Counseling*, vol. 59, no. 1, pp. 21–30, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0738399104003076

[375] B. W. Hesse, D. E. Nelson, G. L. Kreps, R. T. Croyle, N. K. Arora, B. K. Rimer, and K. Viswanath, "Trust and sources of health information: the impact of the internet and its implications for health care providers: findings from the first health information national trends survey," *Archives of internal medicine*, vol. 165, no. 22, pp. 2618–2624, 2005.

[376] A. Broom, "The emale: Prostate cancer, masculinity and online support as a challenge to medical expertise," *J. Sociol. (Melb)*, vol. 41, no. 1, pp. 87–104, Mar. 2005.

[377] S. Kerstan, N. Bienefeld, and G. Grote, "Choosing human over ai doctors? how comparative trust associations and knowledge relate to risk and benefit perceptions of ai in healthcare," *Risk Analysis*, vol. 44, 09 2023.

[378] C. Montag, B. Klugah-Brown, X. Zhou, J. Wernicke, C. Liu, J. Kou, Y. Chen, B. W. Haas, and B. Becker, "Trust toward humans and trust toward artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging," *Personality Neuroscience*, vol. 6, p. e3, 2023.

[379] M. Jakesch, M. French, X. Ma, J. T. Hancock, and M. Naaman, "Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–13. [Online]. Available: https://doi.org/10.1145/3290605.3300469

[380] N. Scharowski, M. Benk, S. J. KÃŒhne, L. Wettstein, and F. BrÃŒhlmann, "Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study," in *2023 ACM Conference on Fairness, Accountability, and Transparency*. Chicago IL USA: ACM, Jun. 2023, pp. 248–260. [Online]. Available: https://dl.acm.org/doi/10.1145/3593013.3593994

[381] M. I. Ahmad, I. Keller, D. A. Robb, and K. S. Lohan, "A framework to estimate cognitive load using physiological data," *Personal Ubiquitous Comput.*, vol. 27, no. 6, p. 2027–2041, sep 2020. [Online]. Available: https://doi.org/10.1007/s00779-020-01455-7

[382] K. Ji, D. Hettiachchi, F. D. Salim, F. Scholer, and D. Spina, "Characterizing information seeking processes with multiple physiological signals," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR 2024, vol. 5. ACM, Jul. 2024, p. 1006–1017. [Online]. Available: http://dx.doi.org/10.1145/3626772.3657793

[383] K. Ji, D. Spina, D. Hettiachchi, F. Scholer, and F. D. Salim, "Towards detecting tonic information processing activities with physiological data," in *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, ser. UbiComp/ISWC '23 Adjunct. New York, NY, USA: Association for Computing Machinery, 2023, p. 1–5. [Online]. Available: https://doi.org/10.1145/3594739.3610679

[384] Y. Abdrabou, E. Karypidou, F. Alt, and M. Hassib, "Investigating user behavior towards fake news on social media using gaze and mouse movements," 01 2023.

[385] Ömer Sümer, E. Bozkir, T. Kübler, S. Grüner, S. Utz, and E. Kasneci, "Fakenewsperception: An eye movement dataset on the perceived believability of news stories," *Data in Brief*, vol. 35, p. 106909, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340921001931

[386] Y. Lu and N. Sarter, "Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability," *IEEE Transactions on Human-Machine Systems*, vol. PP, pp. 1–9, 08 2019.

[387] N. Sevcenko, T. Appel, M. Ninaus, K. Moeller, and P. Gerjets, "Theory-based approach for assessing cognitive load during time-critical resource-managing human–computer interactions: an eye-tracking study," *Journal on Multimodal User Interfaces*, vol. 17, pp. 1–19, 11 2022.

[388] P. Ayres, J. Y. Lee, F. Paas, and J. J. G. van Merriënboer, "The validity of physiological measures to identify differences in intrinsic cognitive load," *Front Psychol*, vol. 12, p. 702538, Sep. 2021.

[389] L. Wang and J. A. Stern, "Saccade initiation and accuracy in gaze shifts are affected by visual stimulus significance," *Psychophysiology*, vol. 38, no. 1, pp. 64–75, Jan. 2001.

[390] L. Wang, "Eye tracking methodology in screen-based usability testing," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–3. [Online]. Available: https://doi.org/10.1145/3290607.3298811

[391] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investig*, vol. 15, no. 3, pp. 235–245, Feb. 2018.

[392] B. Thielmann, J. Hartung, and I. Böckelmann, "Objective assessment of mental stress in individuals with different levels of effort reward imbalance or overcommitment using heart rate variability: a systematic review," *Syst Rev*, vol. 11, no. 1, p. 48, Mar. 2022.

[393] I. B. Ajenaghughrure, S. C. Da Costa Sousa, and D. Lamas, "Psychophysiological modelling of trust in technology: Comparative analysis of algorithm ensemble methods," in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2021, pp. 000 161–000 168.

[394] S. S. Parikh, "Eye gaze feature classification for predicting levels of learning," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53471366

[395] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinform.*, vol. 20, no. 1, pp. 511:1–511:23, 2019. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3119-4

[396] C. D. Norman and H. A. Skinner, "eheals: The ehealth literacy scale," *J Med Internet Res*, vol. 8, no. 4, p. e27, Nov 2006. [Online]. Available: http://www.jmir.org/2006/4/e27/

[397] Prolific, 2014. [Online]. Available: https://www.prolific.com

[398] S. Kherad-Pajouh and O. Renaud, "A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs," *Statistical Papers*, vol. 56, no. 4, pp. 947–967, Nov. 2015.

[399] W. Haynes, *Benjamini–Hochberg Method*. New York, NY: Springer New York, 2013, pp. 78–78. [Online]. Available: https://doi.org/10.1007/978-1-4419-9863-7_1215

[400] T. AB, "Tobii pro lab," Computer software, Danderyd, Stockholm, 2024. [Online]. Available: http://www.tobii.com/

[401] U. van Amsterdam, "Fmg research lab — lab-fmg.uva.nl." [Online]. Available: https://lab-fmg.uva.nl/en

[402] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, no. 1, pp. 195–203, Feb. 2019.

[403] F. Alamudun, J. Choi, R. Gutierrez-Osuna, H. Khan, and B. Ahmed, "Removal of subject-dependent and activity-dependent variation in physiological measures of stress," in *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, 2012, pp. 115–122.

[404] G. Sahagian, "What is random state 42? — grsahagian.medium.com," 2024. [Online]. Available: https://grsahagian.medium.com/what-is-random-state-42

[405] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: A big comparison for nas," 2019. [Online]. Available: https://arxiv.org/abs/1912.06059

[406] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," vol. 14, 03 2001.

[407] J. Opitz and S. Burst, "Macro f1 and macro f1," 2021. [Online]. Available: https://arxiv.org/abs/1911.03347

[408] R. Van der Lans, M. Wedel, and R. Pieters, "Defining eye-fixation sequences across individuals and tasks: the binocular-individual threshold (bit) algorithm," *Behavior research methods*, vol. 43, pp. 239–257, 2011.

[409] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 143–146. [Online]. Available: https://doi.org/10.1145/1978942.1978963

[410] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[411] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[412] S. S. Sundar and J. Kim, "Machine heuristic: When we trust computers more than humans with our personal information," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–9. [Online]. Available: https://doi.org/10.1145/3290605.3300768

[413] C. B. Horton, Jr, M. W. White, and S. S. Iyengar, "Bias against AI art can enhance perceptions of human creativity," *Sci. Rep.*, vol. 13, no. 1, p. 19001, Nov. 2023.

[414] P. Formosa, W. Rogers, Y. Griep, S. Bankins, and D. Richards, "Medical ai and human dignity: Contrasting perceptions of human and artificially intelligent (ai) decision making in diagnostic and medical resource allocation contexts," *Computers in Human Behavior*, vol. 133, p. 107296, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563222001182

[415] Z. Epstein, M. C. Fang, A. A. Arechar, and D. G. Rand, "What label should be applied to content produced by generative ai?" Jul 2023. [Online]. Available: osf.io/preprints/psyarxiv/v4mfz

[416] J. De Freitas, S. Agarwal, B. Schmitt, and N. Haslam, "Psychological factors underlying attitudes toward ai tools," *Nat Hum Behav*, vol. 7, no. 11, pp. 1845–1854, Nov. 2023.

[417] A. E. Scott, D. Neumann, J. Niess, and P. W. Woźniak, "Do you mind? user perceptions of machine consciousness," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3581296

[418] O. Burrus, A. Curtis, and L. Herman, "Unmasking ai: Informing authenticity decisions by labeling ai-generated content," *Interactions*, vol. 31, no. 4, p. 38–42, jun 2024. [Online]. Available: https://doi.org/10.1145/3665321

[419] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*, vol. 66, no. 11, pp. 4944–4957, 2020. [Online]. Available: https://doi.org/10.1287/mnsc.2019.3478

[420] M. A. Just and P. A. Carpenter, "A theory of reading: From eye fixations to comprehension," *Psychol. Rev.*, vol. 87, no. 4, pp. 329–354, 1980.

[421] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychol. Bull.*, vol. 124, no. 3, pp. 372–422, 1998.

[422] N. Boonprakong, X. Chen, C. Davey, B. Tag, and T. Dingler, “Bias-aware systems: Exploring indicators for the occurrences of cognitive biases when facing different opinions,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3580917

[423] Q. V. Liao, M. Vorvoreanu, H. Subramonyam, and L. Wilcox, “Ux matters: The critical role of ux in responsible ai,” *Interactions*, vol. 31, no. 4, p. 22–27, jun 2024. [Online]. Available: https://doi.org/10.1145/3665504

[424] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna, “Explanations can reduce overreliance on ai systems during decision-making,” *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW1, apr 2023. [Online]. Available: https://doi.org/10.1145/3579605

[425] J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, *Strong Inference in Psychophysiological Science*, ser. Cambridge Handbooks in Psychology. Cambridge University Press, 2016, p. 3–15.

[426] D. Michalkova, M. P. Rodriguez, and Y. Moshfeghi, “Understanding feeling-of-knowing in information search: An eeg study,” *ACM Trans. Inf. Syst.*, vol. 42, no. 3, jan 2024. [Online]. Available: https://doi.org/10.1145/3611384

[427] J. T. Cacioppo, P. A. Rourke, B. S. Marshall-Goodell, L. G. Tassinary, and R. S. Baron, “Rudimentary physiological effects of mere observation,” *Psychophysiology*, vol. 27, no. 2, pp. 177–186, 1990. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1990.tb00368.x

[428] R. D. Crosby, J. M. Lavender, S. G. Engel, and S. A. Wonderlich, *Ecological Momentary Assessment*. Singapore: Springer Singapore, 2016, pp. 1–3. [Online]. Available: https://doi.org/10.1007/978-981-287-087-2_159-1

[429] M. S. Mounica, M. Manvita, C. Jyotsna, and J. Amudha, “Low cost eye gaze tracker using web camera,” in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019, pp. 79–85.

[430] D. J. McDuff, S. Gontarek, and R. W. Picard, “Remote measurement of cognitive stress via heart rate variability,” *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2957–2960, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:206627980

[431] E. A. Act, “Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act) (text with eea relevance),” https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024, [Accessed 12-09-2024]. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj

[432] S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke, “How weird is chi?” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445488

[433] “Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance),” pp. 1–88, May 2016.

[434] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218305988

[435] X. Yin and X. Wan, “How do Seq2Seq models perform on end-to-end data-to-text generation?” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7701–7710. [Online]. Available: https://aclanthology.org/2022.acl-long.531/

[436] C. J. Robins, N. Zerubavel, A. M. Ivanoff, and M. M. Linehan, “Dialectical behavior therapy,” in *Handbook of personality disorders: Theory, research, and treatment*, W. J. Livesley and R. Larstone, Eds. The Guilford Press, 2018, pp. 527–540.

[437] M. A. Manzoor, Y. Wang, M. Wang, and P. Nakov, “Can machines resonate with humans? evaluating the emotional and empathic comprehension of LMs,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 14 683–14 701. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.861/

[438] K. Meding, “It’s complicated. the relationship of algorithmic fairness and non-discrimination regulations in the eu ai act,” 2025. [Online]. Available: https://arxiv.org/abs/2501.12962

[439] X. Wang, Y. Yang, D. Tao, and T. Zhang, “The impact of ai transparency and reliability on human-ai collaborative decision-making,” in *AHFE International*. AHFE International, 2023.

[440] S. Desai, C. Z. Wei, J. Sin, M. Dubiel, N. Zargham, S. Ahire, M. Porcheron, A. Kuzminykh, M. Lee, H. Candello, J. E. Fischer, C. Munteanu, and B. R. Cowan, “Cui@chi 2024: Building trust in cuis—from design to deployment,” in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613905.3636287

[441] J. Li, H. Cao, L. Lin, Y. Hou, R. Zhu, and A. El Ali, “User experience design professionals’ perceptions of generative artificial intelligence,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI ’24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642114

[442] N.-N. Zhang, S.-Z. Ye, and T.-Y. Chien, “Imbalanced data classification based on hybrid methods,” in *Proceedings of the 2nd International Conference on Big Data Research*, ser. ICBDR ’18. New York, NY, USA: Association for Computing Machinery, 2018, p. 16–20. [Online]. Available: https://doi.org/10.1145/3291801.3291812

[443] J. Passmore, “Motivational interviewing techniques reflective listening.”

Summary

In this thesis, we explore two interrelated themes in the application of large language models (LLMs) in health contexts: (1) aligning LLMs with domain expertise (i.e., expert-crafted dialogue scripts and therapeutic strategies) in delivering digital psychotherapy for health intervention, to enhance controllability and explainability; and (2) understanding people's trust perception in LLM-powered health information seeking. Through a combination of computational modeling, empirical studies, and theoretical analysis, this work advances both the methodological alignment of LLMs with therapeutic principles and the human-centered evaluation, as well as the design of LLM-driven health communication.

The first theme of this thesis focuses on enhancing the controllability and explainability of LLM-powered psychotherapeutic health interventions by aligning them with domain expertise. ***Chapter 2*** initially investigates how well LLMs can generate reflective utterances, a core therapist's behavior in MI, across three prominent LLMs (GPT-4, Llama-3, and BLOOM). Evaluations using both automatic and human assessments reveal that while LLMs demonstrate promising capabilities in generating MI reflections, they still face challenges in producing deeper, emotionally nuanced, and contextually adaptive dialogue responses. ***Chapter 3*** addresses the challenge of data scarcity in the area of psychotherapy by creating BiMISC, a bilingual MI dataset with expert annotations following the Motivational Interviewing Skills Code (MISC) scheme [11]. As a valuable resource for the subsequent work in this thesis on aligning language models for psychotherapeutic dialogue generations. Expanding on this, ***Chapter 4*** presents an expertise-driven alignment approach (i.e., Script-Strategy Aligned Generation, SSAG), where LLMs first predict MI strategies before generating therapeutic dialogue responses. This structured approach improves explainability and dialogue adherence to therapeutic principles, offering a controllable way to integrate LLMs into psychotherapy applications. Experimental results confirm that aligning LLMs with domain expertise can enhance both adherence and therapeutic appropriateness, providing a foundation for future LLM-assisted counseling tools for

health behavior intervention.

The second theme of the thesis examines trust in LLM-powered health information seeking, focusing on how search agents and dissemination interfaces influence people's trust perception in online personal health information. ***Chapter 5*** investigates how trust varies between traditional search engines (i.e., Google) and LLM-powered conversational search (i.e., ChatGPT). Results indicate that participants trust ChatGPT's health information more than Google's due to its conversational and personalized responses, yet longstanding trust in traditional search engines still plays a role. The research further explores how different dissemination interfaces (i.e., text-based, speech-based, and embodied) affect trust perceptions in LLM-generated information provided, revealing that participants prefer text-based interactions for self-verifying information, while speech-based and embodied interfaces enhance engagement but introduce additional trust-related concerns. ***Chapter 6*** extends this research by investigating how source attribution and transparency labeling (AI-generated vs. human-generated) impact trust in online health information. Using a mixed-methods study with behavioral (i.e., eye-tracking) and physiological sensing (i.e., ECG, EDA, and body temperature), the study finds that LLM-generated health information is generally trusted more than human-written content. However, information labeled as human-sourced is perceived as more trustworthy than information labeled as AI-generated, regardless of actual source. These findings underscore the nuanced relationship between transparency of source and disclosed labeling and cognitive trust calibration in AI-powered health communication.

Taken together, this thesis advances both the development of LLM-driven psychotherapy applications for health behavior intervention and the understanding of trust in LLM-powered health information seeking. The research in this thesis highlights the importance of alignment approaches for enhancing LLMs' explainability and controllability in sensitive therapeutic contexts from a computational perspective while also emphasizing the crucial role of user interface design and transparency mechanisms in shaping trust in digital health applications through human-centric insights. By bridging technical advancements with empirical insights into human-AI interaction, this work lays the foundation for the explainable and transparent deployment of LLMs in real-world health contexts.

Nederlandse samenvatting

In dit proefschrift verkennen wij twee onderling verbonden thema's in de toepassing van large language models (LLMs) in gezondheidscontexten: (1) het afstemmen van LLMs op domeinexpertise (d.w.z. door experts ontworpen dialoogscripts en therapeutische strategieën) bij het aanbieden van digitale psychotherapie voor gezondheidsinterventie, om de controleerbaarheid en verklaarbaarheid te vergroten; en (2) het begrijpen van hoe mensen vertrouwen ervaren in LLM-gestuurde zoekprocessen naar gezondheidsinformatie. Door een combinatie van computationele modellering, empirische studies en theoretische analyse draagt dit werk bij aan zowel de methodologische afstemming van LLMs op therapeutische principes als aan de mensgerichte evaluatie en het ontwerp van LLM-gedreven gezondheidscommunicatie.

Het eerste thema van deze thesis richt zich op het versterken van de controleerbaarheid en verklaarbaarheid van LLM-gestuurde psychotherapeutische gezondheidsinterventies door ze af te stemmen op domeinexpertise. ***Hoofdstuk 2*** vormt een eerste stap en onderzoekt hoe goed LLMs reflectieve uitingen kunnen genereren – een kernvaardigheid van therapeuten binnen Motiverende Gespreksvoering (MI) – met behulp van drie prominente LLMs (GPT-4, Llama-3 en BLOOM). Evaluaties met zowel automatische als menselijke beoordelingen laten zien dat LLMs veelbelovende capaciteiten tonen in het genereren van MI-reflecties, maar nog steeds moeite hebben met het produceren van diepere, emotioneel genuanceerde en contextueel adaptieve dialoogresponsen. ***Hoofdstuk 3*** pakt de uitdaging van dataschaarste in psychotherapie aan door BiMISC te creëren, een tweetalige MI-dataset met expertannotaties volgens het Motivational Interviewing Skills Code (MISC)-schema [11]. Deze dataset vormt een waardevolle bron voor het vervolgonderzoek in dit proefschrift naar het afstemmen van taalmodellen voor psychotherapeutische dialooggeneratie. Voortbouwend hierop presenteert ***Hoofdstuk 4*** een expertisegestuurde afstemmingsaanpak (d.w.z. Script-Strategy Aligned Generation, SSAG), waarbij LLMs eerst MI-strategieën voorspellen alvorens therapeutische dialoogresponsen te genereren.

Deze gestructureerde aanpak verbetert de verklaarbaarheid en de naleving van therapeutische principes, en biedt een gecontroleerde manier om LLMs te integreren in psychotherapie. Experimentele resultaten bevestigen dat afstemming op domeinexpertise zowel de naleving als de therapeutische geschiktheid kan vergroten, en zo een basis legt voor toekomstige LLM-ondersteunde counselingtools voor gezondheidsinterventie.

Het tweede thema van dit proefschrift onderzoekt vertrouwen in LLM-gestuurde zoekprocessen naar gezondheidsinformatie, met nadruk op hoe zoekagenten en hun interfaces de perceptie van vertrouwen beïnvloeden bij online persoonlijke gezondheidsinformatie. ***Hoofdstuk 5*** onderzoekt hoe vertrouwen verschilt tussen traditionele zoekmachines (zoals Google) en LLM-gestuurde conversationele zoekoplossingen (zoals ChatGPT). De resultaten tonen aan dat deelnemers de gezondheidsinformatie van ChatGPT meer vertrouwen dan die van Google vanwege de conversationele en gepersonaliseerde antwoorden, al speelt het langdurige vertrouwen in traditionele zoekmachines nog steeds een rol. In het onderzoek wordt verder verkend hoe verschillende interfaces (d.w.z. tekstgebaseerd, spraakgebaseerd en belichaamd in de vorm van een robot) het vertrouwen in LLM-gegenereerde informatie beïnvloeden. Hieruit blijkt dat deelnemers de voorkeur geven aan tekstgebaseerde interacties om informatie zelfstandig te verifiëren, terwijl spraakgebaseerde en belichaamde interfaces de betrokkenheid vergroten, maar ook extra vertrouwensgerelateerde zorgen oproepen. ***Hoofdstuk 6*** breidt dit onderzoek uit door te analyseren hoe bronattributie en transparantielabeling (AI-gegenereerd versus menselijk-gegenereerd) vertrouwen in online gezondheidsinformatie beïnvloeden. Met behulp van een *mixed-methods* benadering, waaronder gedragsmetingen (zoals eye-tracking) en fysiologische sensoren (zoals ECG, EDA en lichaamstemperatuur), laat de studie zien dat LLM-gegenereerde gezondheidsinformatie over het algemeen meer wordt vertrouwd dan door mensen geschreven inhoud. Toch wordt informatie die als menselijk afkomstig wordt gelabeld betrouwbaarder gevonden dan informatie die als AI-gegenereerd is gelabeld, ongeacht de werkelijke bron. Deze bevindingen benadrukken de genuanceerde relatie tussen transparantie, bronvermelding en cognitieve kalibratie van vertrouwen in AI-gestuurde gezondheidscommunicatie.

Samengevat draagt dit proefschrift bij aan zowel de ontwikkeling van LLM-gedreven psychotherapieapplicaties voor gezondheidsinterventies als aan het begrip van vertrouwen in LLM-gestuurde gezondheidsinformatievoorziening. Het onderzoek in dit proefschrift benadrukt het belang van afstemmingsmethoden om de verklaarbaarheid en controleerbaarheid van LLMs in gevoelige therapeutische contexten te versterken vanuit een computationeel perspectief, en legt tevens de nadruk op de cruciale rol van interfaceontwerp en transparantiemechanismen in het vormen van vertrouwen in digitale gezondheidsapplicaties vanuit een mensgericht perspectief. Door technische vooruitgang te verbinden met empirische inzichten in mens-AI-interactie, legt dit werk de basis voor een verklaarbare en transparante inzet van LLMs in realistische gezondheidscontexten.

Disclosure of Artificial Intelligence (AI) Usage

In preparing this thesis, GenAI tools were used for correcting and editorially improving texts to enhance language use, including spelling and grammar, paraphrasing, and improving style, readability, and conciseness of sentences and paragraphs.

For this purpose, I made use of ChatGPT (model GPT-4o, publicly available web version, not the API). I remain fully responsible for the intellectual content of this thesis, and the use of GenAI was in line with both the “Policy Framework and Guidelines on GenAI in Education” at the University of Amsterdam [1] as well as the “Preliminary FMG guidelines Integrity and transparency of use Generative AI in research” from the Faculty of Social and Behavioural Sciences at the University of Amsterdam.

By submitting this thesis for assessment, I thus confirm that the thesis is my own intellectual property, and that the ideas as well as text from other sources have been properly cited. All quotes and sourced information are identifiable as such. With this statement, I have disclosed all AI technology and its use in that supported the writing process.

[1] Universiteit van Amsterdam. (2025, May). *Policy framework and guidelines on GenAI in education.* https://www.uva.nl/en/about-the-uva/policy-and-regulations/education/policy-framework-and-guidelines-on-genai-in-education.html

Acknowledgements

As this four-year chapter draws to a close, I am filled with deep gratitude. I have never been someone good with words, nor one easily showing emotions, but over these years, I have been truly lucky to meet so many wonderful people. This journey would not have been possible without those who walked beside me, each leaving a lasting mark in their own way. To everyone sharing a part of this journey, thank you. When you think of me, I hope you know how deeply grateful I am for your companionship, whether it was for a stop, a stage, or the entire way.

First and foremost, my deepest gratitude goes to my supervisors, Dr Jos A. Bosch, Dr Abdallah El Ali, and Dr Jan de Wit, the architects of my growth as a researcher. Jos, thank you for your trust, freedom, and unwavering support throughout my PhD. You gave me the confidence to explore what truly fascinated me. Your calm confidence and ability to handle complexity with grace and peace have deeply inspired me. Abdo, thank you for being a generous and thoughtful mentor. You taught me to see the human side of technology. I am especially grateful for your patience with my countless "quick questions", your honest feedback, and many deep conversations that expanded my thinking and comforted me during difficult times. I feel very fortunate to have met and learned from you during my PhD journey. Jan, your insightful feedback always went straight to the heart of the matter, and your warm words never failed to lift my spirits when challenges felt heavy. I am truly grateful to have known you and you have always been a source of warmth and encouragement.

To my colleagues at UvA and CWI, thank you for creating an environment full of curiosity, generosity, and shared purpose. Working with you made research not just an individual pursuit, but a collaborative and joyful experience. From coffee chats to endless deadlines, your kindness and encouragement made this journey lighter and more meaningful. I am also deeply thankful to my collaborators, many of whom have become not only research partners but also dear friends in my life. A special thanks goes to the TIMELY project, which supported my PhD. I am grateful to everyone involved in TIMELY for the inspiring discussions and interdisciplinary collaborations that broadened my perspective and enriched my research.

To all my friends, near and far, thank you for being my anchor beyond academia and for bringing so much warmth into my life, especially during the challenging time to get the PhD. You reminded me, again and again, that there is more to life than papers and deadlines. Thank you for every shared meal, long walk, late-night talk, and moment of sport that made the hard days lighter. Whether we were exploring the cities or simply sending messages across time zones, your presence, near or distant, kept me grounded, joyful, and hopeful. This journey would have been far lonelier without you, and I am deeply grateful for your friendship and care.

And above all, to my parents and family, my unwavering foundation. Your unconditional love and belief gave me the strength to persevere through every challenge. I owe so much to my mom, whose trust and support have been indispensable to this journey. Lastly, my heartfelt thanks to Chengcheng, whose companionship has been a constant source of strength and calm throughout the years.

“你不必生来勇敢，天赋过人，只要能投入勤奋，诚诚恳恳。”

八年前，我带着这句话，踏上了探索欧罗巴的旅途，也开启了自我探索的过程。当时未曾想，这趟旅程一走，竟是八年时光。

德国的三年，是在海德堡的宁静中慢慢沉淀下的时光。那段日子让我习惯了独立，也逐渐找到了与自己相处的方式，在学习与思考中积蓄前行的力量。毕业之际，我在王座山顶写下这句话：“感恩能在这样的小城，度过三年时光：宁静祥和，却充实奋进。感谢所有曾经、现在，支持、理解和陪伴我的人，也要感谢在20多岁的路上，经历过的所有好与不好。愿自己在人生新的征途上，一切顺遂。”至今读来，依然承载着我的初心、感激与期翼。

四年后的今天，站在阿姆斯特丹大学的礼堂，我依然想说：感恩。阿姆斯特丹不似海德堡的宁静，却以另一种方式，让我在它的多元与繁华中砥砺，让我学会在人群中独立思考，在纷繁中坚定自己。这段旅程，将这份“充实奋进”打磨得更加坚韧。这篇博士论文，便是这八年时光下的一份答卷。它不仅是一段学业的终点，更是对那句“投入勤奋，诚诚恳恳”的漫长践行与回响。

我感恩命运的馈赠，让我得以远行，看到更广阔的世界，也更有幸能在学术的道路上留下属于自己的印记。这条漫漫之路，若无师长亲友的同行与扶持，不过一场孤独的漂泊。正是这些温暖的相遇，为我点亮了夜航时的星图。这份感恩，难以尽述，唯有铭记于心。

木欣欣以向荣，泉涓涓而始流。此去前路，愿有光，亦有星辰。

Publications

Papers in Conference Proceedings (Peer Reviewed)

1. ***Xin Sun***, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, Jos Bosch. Rethinking the Alignment of Psychotherapy Dialogue Generation with Motivational Interviewing Strategies. (Oral) In Proceedings of the International Conference on Computational Linguistics (COLING 2025)

2. ***Xin Sun****, Erkan Basar*, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A.Bosch, Emiel Krahmer. How Well Can LLMs Reflect? A Human Evaluation of LLM-generated Reflections for Motivational Interviewing Dialogues. (Oral) In Proceedings of the International Conference on Computational Linguistics (COLING 2025) (*equal contribution.)

3. ***Xin Sun***, Lei Wang, Yue Li, Jie Li, Massimo Poesio, Julian Frommel, Koen Hindriks, Jiahuan Pei. Talking-to-Build: How LLM-Assisted Interface Shapes Player Performance and Experience in Minecraft. ACM International Conference on Multimodal Interaction (ICMI 2025)

4. ***Xin Sun***, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos TP Dobber, Jos Bosch. Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)

5. ***Xin Sun***, Isabelle Teljeur, Zhuying Li, Jos A. Bosch. Can a Funny Chatbot Make a Difference? Infusing Humor into Conversational Agent for Behavioral Intervention. ACM Conference on Conversational User Interfaces (CUI 2024)

6. ***Xin Sun***, Yunjie Liu, Jan De Wit, Jos A. Bosch, and Zhuying Li. Trust by Interface: How Different User Interfaces Shape Human Trust in Health Information from Large Language Model. (Late-Breaking Work) ACM Conference on Human Factors in Computing Systems (CHI 2024)

7. ***Xin Sun***, Emiel Krahmer, Jan De Wit, Reinout Wiers, and Jos A. Bosch. Plug and Play Conversations: The MicroConversation Scheme for Modular Development of Hybrid CA. (Poster) ACM SIGCHI Conference on Computer-Supported Cooperative Work And Social Computing (CSCW 2023)

8. ***Xin Sun***, Jos A. Bosch, Jan De Wit, and Emiel Krahmer. Human-in-the-Loop Interaction for Continuously Improving Generative Model in CAs for Behavioral Intervention. (Poster) ACM Conference on Intelligent User Interfaces (IUI 2023)

9. ***Xin Sun***, Boris Schmitz, and Jos A. Bosch. TIMELY: Providing In-Time Support for Cardiovascular Rehabilitation with “Patients and Practitioners in the Loop”. (Poster) ACM Conference on Intelligent User Interfaces (IUI 2023)

10. ***Xin Sun***, et al. Virtual Support for Real-World Movement: Using Chatbots to Overcome Barriers to Physical Activity. Frontiers in Artificial Intelligence and Applications (HHAI 2023)

11. Jiahuan Pei, Fanghua Ye, ***Xin Sun***, Wentao Deng, Koen Hindriks, Junxiao Wang. Conversational education at scale: A multi-llm agent workflow for procedural learning and pedagogic quality assessment. The Conference on Empirical Methods in Natural Language Processing (EMNLP 2025).

12. Yuto Mandai, Katie Seaborn, Tomoyasu Nakano, ***Xin Sun***, Yijia Wang, Jun Kato. Super Kawaii Vocalics: Amplifying the “Cute” Factor in Computer Voice. ACM Conference on Human Factors in Computing Systems (CHI 2025).

13. Xiao Tang, Zhuying Li, ***Xin Sun***, Xuhai ”Orson” Xu, Min-Ling Zhang. ZzzMate: Designing an Empathetic Chatbot for Addressing Self-Conscious Emotions for Sleep Adherence. (Late-Breaking Work) ACM Conference on Human Factors in Computing Systems (CHI 2025).

14. Xiao Wang, Jiahuan Pei, Diancheng Shui, Zhiguang Han, ***Xin Sun***, Dawei Zhu,

Xiaoyu Shen. MultiJustice: A Chinese Dataset for Multi-Party, Multi-Charge Legal Prediction. CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC 2025).

15. Zhuying Li, Si Cheng, Zhenhuan Chen, ***Xin Sun***, Jiatong Li, Ding Ding. SleepyFlora: Supporting Sleep Sharing and Augmentation over a Distance for Social Bonding across Time Zones. (Poster) ACM SIGCHI Conference on Computer-Supported Cooperative Work And Social Computing (CSCW 2023)

16. Benjamin Beilharz, ***Xin Sun***, Sariya Karimova, and Stefan Riezler. LibriVoxDeEn-A Corpus for German-to-English Speech Translation and Speech Recognition. International Conference on Language Resources and Evaluation (LREC 2020)

Journal and Book

1. ***Xin Sun***, Jan de Wit, Zhuying Li, Jiahuan Pei, Abdallah El Ali, Jos A Bosch. Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy. The Proceedings of the ACM on Human Computer Interaction (Journal PACMHCI) Track Computer-Supported Cooperative Work and Social Computing (CSCW)

2. ***Xin Sun***, Yunjie Liu, Jos A. Bosch, and Zhuying Li. Interface Matters: Exploring Trust Perception in Health Information from Large Language Models via Text, Speech, and Embodiment. The Proceedings of the ACM on Human Computer Interaction (Journal PACMHCI) Track Computer-Supported Cooperative Work and Social Computing (CSCW)

3. ***Xin Sun***, Rongjun Ma, Shu Wei, Pablo César Garcia, Jos A. Bosch, Abdallah El Ali. Understanding Trust Perception in AI-Generated Information with Behavioral and Physiological Sensing. Under review by the International Journal of Human-Computer Studies. (Major revision)

4. ***Xin Sun***, Rongjun Ma, Xiaochang Zhao, Janne Lindqvist, Jan de Wit, Zhuying Li, Abdallah El Ali, Jos A.Bosch. From Agents to Interfaces: Understanding Trust in Health Information from Conversational Search. Under review by the Behaviour and Information Technology Journal. (Major revision)

5. Xiao Tang, Zhuying Li, ***Xin Sun***, Xuhai "Orson" Xu, Min-Ling Zhang. ZzzMate: Designing an Empathetic Chatbot for Addressing Self-Conscious Emotions for Sleep Adherence. International Journal of Human-Computer Interaction.

6. Zhuying Li, Yan Wang, ***Xin Sun***. Integrating Culture in Human-Food Interaction: A Study of Cultural and Creative Food Experiences and Technological Interactions. International Journal of Human-Computer Studies.

7. Mirela Habibovic, Emma Douma, Hendrik Schäfer, Manuela Sestayo-Fernandez, Tom Roovers, ***Xin Sun***, ... Boris Schmitz. A patient-centered risk prediction, prevention and intervention platform to support the continuum of care in coronary disease using eHealth and artificial intelligence. Journal of Medical Internet Research.

8. Gert-Jan de Bruijn, Divyaa Balaji and ***Xin Sun***. Book Publication: "Chapter 16

Chatbots for Health Communication". Book: Health, Media, and Communication, De Gruyter Mouton.

Curriculum Vitae

Xin Sun

Amsterdam, the Netherlands

Email: xin.von.sun@gmail.com | **Web:** https://xin-von-sun.github.io/

EDUCATION

PhD: University of Amsterdam, the Netherlands *2021 – 2025*
Thesis: From Aligned Models to Trusted Interfaces: Explainable Health Intervention and Transparent Health Information Seeking

MSc: Heidelberg University, Germany *2017 – 2020*
Major: Scientific Computing
(Mathematics & Natural Language Processing)

BSc: Xidian University, China *2011 – 2015*
Major: Computational Mathematics

EXPERIENCE

Centrum Wiskunde & Informatica (CWI), the Netherlands *2023 – 2025*
PhD Researcher

Alibaba Group, China *2020*
NLP Algorithm Engineer *(Full-time Internship)*

SAP, Germany *2019*
Data Scientist *(Full-time Internship)*