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

Xin Sun[*†], **Jiahuan Pei**[†§], **Jan de Wit**[‡], **Mohammad Aliannejadi**[*], **Emiel Krahmer**[‡], **Jos T.P. Dobber**[‖], **and Jos A. Bosch**[*]

[*]University of Amsterdam
Amsterdam, The Netherlands
{x.sun2, m.aliannejadi, j.a.bosch}@uva.nl

[†]Centrum Wiskunde & Informatica (CWI)
Amsterdam, The Netherlands
j.pei@uva.nl

[‡]Tilburg University
Tilburg, The Netherlands
{j.m.s.dewit, e.j.krahmer}@tilburguniversity.edu

[‖]Hogeschool van Amsterdam
Amsterdam, The Netherlands
j.t.p.dobber@hva.nl

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 introduce 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 present 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 demonstrate 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.

Keywords: Bilingual Motivational Interviewing dataset, Large language model for Motivational Interviewing coding, Low-resourced languages

1. Introduction

Motivational interviewing (MI) is an essential, directive, client-centered counseling technique, that aims to elicit clients' behavioral change (Miller and Rollnick, 2002). It can boost intrinsic motivation and collaboration between therapists and clients by effectively addressing ambivalence and enhancing self-efficacy (Martins and McNeil, 2009), improving clients' adherence to the therapists' interventions (Alperstein and Sharpe, 2016). Without the use of MI, traditional techniques can potentially cause resistance and disengagement from clients due to their confrontational, paternalistic ways of thinking (Miller and Rollnick, 2002).

Behavioral coding (BC) is the practice of systematically observing and categorizing the behaviors of therapists and clients (Martins and McNeil, 2009; Miller and Rollnick, 2002; Tavabi et al., 2021). 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 BC in MI is challenging, as it relies on the expertise (Jannet M. de Jonge, 2005; Atkins et al., 2014) and large-scale datasets with human annotations (Cao et al., 2019; Tanana et al., 2016; Klonek et al., 2015). Recent research has demonstrated the effectiveness of natural language processing (NLP) approaches in BC. Early efforts primarily utilized statistical models such as N-grams (Pérez-Rosas et al., 2017) and CRFs (Can et al., 2015). More recent approaches have shifted towards deep learning models, including RNN (Tavabi et al., 2021; Xiao et al., 2016), CNN (Wu et al., 2023), and BiGRU with attention mechanism (Cao et al., 2019).

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

Table 1: An example of annotation of a client utterance in BiMISC dataset and AnnoMI dataset. The use of fine-grained multiple codes allows for a multi-perspective and in-depth understanding of the client's intentions. These are non-trivial insights for therapists to conduct MI treatment in psychotherapy.

[§]Corresponding author

LREC-COLING 2024, pages 5609–5621
20-25 May, 2024.

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, Pérez-Rosas et al. (2016) 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 (Zixiu et al., 2022) and MITI (Welivita and Pu, 2022). However, there are still challenges: (1) they do not consist of conversations from the real MI counseling sessions; (2) AnnoMI comprises only six coarse-grained motivational interviewing skills code (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 introduce motivational interviewing skills code (MISC) scheme for behavioral coding and create BiMISC, a bilingual dataset both in English and Dutch: (1) BiMISC consists conversations collected from real MI counseling sessions in psychotherapy; (2) BiMISC comes with fine-grained behavioral codes strictly grounded on 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 (Deng et al., 2023b) and eliciting expertise in various domain-specific applications, e.g., medical treatment (Yan et al., 2022; **?**), legal judgment analysis (Deng et al., 2023a) and qualitative data analysis (Chew et al., 2023; Paoli, 2023; Tai et al., 2023). In this work, we aim to explore the potential of utilizing the MISC scheme with LLMs to directly generate MISC codes that closely align with expert annotations. Specifically, we leverage MISC codes and their definitions in the MISC manual to elicit expertise from an LLM for MI coding. We first design a prompt template, including task instruction, MISC manual, MISC examples, and historical conversations. Next, we randomly select approximately 3% of the data to serve as test samples for conducting trial experiments on MISC coding. We continuously refine the prompt manually until the output codes are in alignment with human annotations. Last, we conduct experiments and evaluations on the full dataset.

Our contributions can be summarized as follows:

- We collect and release BiMISC, the first bilingual motivational interviewing dataset in both English and Dutch with expert-annotated MISC codes;
- We propose a MISC coding approach by eliciting MISC expertise from LLMs;
- We conduct extensive experiments and analysis of the proposed approach, demonstrating its effectiveness in MI behavioral coding.

2. Related Work

2.1 Motivational interviewing

Motivational interviewing (MI) (Miller and Rollnick, 2002) 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 (Martins and McNeil, 2009). 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 (Alperstein and Sharpe, 2016).

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 (Miller and Rollnick, 2002; Tavabi et al., 2021). 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 (Jannet M. de Jonge, 2005; Martins and McNeil, 2009). 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 BC in MI. Notable among these are Motivational interviewing skills code (MISC) (Miller et al., 2002; Jannet M. de Jonge, 2005) and Motivational interviewing treatment integrity (MITI) (Moyers et al., 2016). 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 opt for MISC coding scheme (Miller et al., 2002) to annotate behsviors of both therapist and client.

Figure 1: Process of the construction of the bilingual dataset: BiMISC.

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 (Street, 2023) are not publicly accessible because of privacy. While annotated MI datasets exist, such as the collection of MI conversational recordings by (Pérez-Rosas et al., 2016), these data are not publicly accessible. To the best of our knowledge, there are two publicly available MI datasets. The first one is AnnoMI (Zixiu et al., 2022), 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 the subset/regroup of MISC). The second one MI dataset (Welivita and Pu, 2022) comprises dialogues from social forums. These dialogues are annotated based on the MITI (Moyers et al., 2016) coding scheme by crowdsourcing annotators. In this work, we introduce 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 is annotated strictly grounded on motivational interviewing skills code (MISC) scheme (Miller et al., 2002) by MI experts.

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 (Jannet M. de Jonge, 2005; Atkins et al., 2014). 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 (Pérez-Rosas et al., 2017), topic models (Atkins et al., 2014), and CRFs (Can et al., 2015) for MI coding. With advancements in computational power, the focus has shifted towards deep learning models. Recent work has studied the applications of RNNs (Tavabi et al., 2021; Xiao et al., 2016), CNNs (Wu et al., 2023), and BiGRUs with attention mechanisms (Cao et al., 2019). 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 (Deng et al., 2023b) and in showcasing expertise across various domain-specific applications, such as medical treatment (Yan et al., 2022; Korngiebel and Mooney, 2021), legal judgment analysis (Deng et al., 2023a), and qualitative data analysis (Xiao et al., 2023; Paoli, 2023; Tai et al., 2023), especially in zero-shot scenarios (Brown et al., 2020; Gao et al., 2021). Given their capabilities, LLMs offers potential for MI coding, which could alleviate the need for extensive training data required by previous research (Tavabi et al., 2021; Xiao et al., 2016; Cao et al., 2019; Wu et al., 2023). Therefore, we explore the feasibility of eliciting domain expertise of MISC scheme from LLMs for efficient MISC coding.

3. Dataset Creation

In this section, we outline the creation of the BiMISC dataset. First, we collect raw conversations between therapists and clients from MI counseling sessions (§ 3.1). Second, we introduce MISC scheme for human annotation (§ 3.2). Last, we report the statistics of the proposed BiMISC dataset (§ 3.3).

3.1 Raw data collection

Initially, we collect 80 audio recordings of conversations between 18 clients and therapists in real MI counseling sessions, conducted in Dutch. Secondly, the therapists transcribe the recordings of 80 Dutch conversations, each averaging 108 utterances. The MI therapists rectify typos and grammar errors, while also anonymizing sensitive information (e.g., names and addresses) in the transcripts. Next, we utilize a machine translator (Google, 2023) to translate the Dutch conversations into English and engage two Dutch Master's students to post-edit the translations. Our raw conversations have significantly more turns, making the utterance count comparable to AnnoMI (average of 108 vs. 80 respectively). Moreover, our MI conversations are from real counseling sessions, ensuring both authenticity and relevance.

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 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.

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

3.2 MISC annotation

We follow the MISC 2.1 scheme (Miller et al., 2002)* and define an annotation manual that contains MISC codes with their descriptions and examples (See Table 2). The certified MI therapists from the Dutch institute who conducted the MI counselings and initially recorded the conversation assign each utterance with the appropriate MISC codes. Each utterance is coded to reflect all applicable MISC behaviors, leading to a multi-code annotation. For example, as shown in Table 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 annotate the most precise fine-grained MISC codes, as shown in Table 2. To ensure the quality of the dataset and its annotations, we provide annotators with a comprehensive guideline. They are encouraged to flag and discuss ambiguous cases. Furthermore, a subset of the data undergoes double-annotation to assess inter-annotator agreement, ensuring the reliability and consistency of the annotations.

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

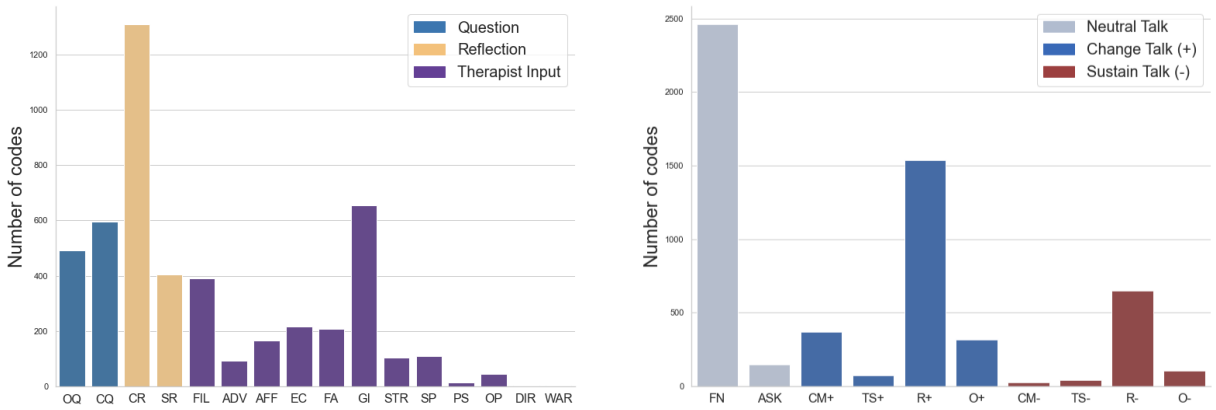

Figure 2: Distribution of fine-grained codes for the client (right) and therapist (left) in BiMISC.

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

Table 4: Comparison of 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 utilization of fine-grained multiple codes provides therapists with profound insights for conducting MI treatment in psychotherapy.

3.3 Data statistics

Table 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 2).

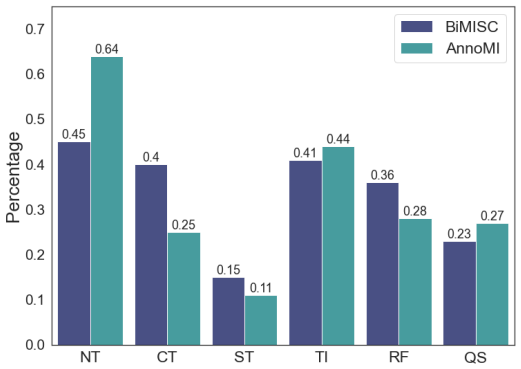

Figure 3: Distribution of coarse-grained codes in AnnoMI dataset and BiMISC dataset.

Table 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 (CQ)" can be mapped to the coarse-grained code "question (QS)".

Figure 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. We see that BiMISC has a more balanced distribution of the codes compared with AnnoMI.

4. Experimental Setup

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?

4.2 Task definition and evaluation

We define the MISC coding task as the classification of therapist or client utterances into specific MISC codes. Utilizing an LLM, we provide 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 conduct evaluation 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 (Opitz and Burst, 2021): provides a well-rounded metric that factors in both precision and recall. This is important given the imbalanced distribution of codes in MI conversations.

4.3 Benchmark models

We employ three prominent LLMs as benchmarks, including two commercial and one open-source.

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

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

Flan-T5 We select `flan-t5-xxl` [§] as an open-source LLM benchmark, known for its advanced capabilities in various NLP tasks (Chung et al., 2022), optimized for better alignment with human instructions.

We also explored `Llama-2-13b-chat-hf` [¶] as an open-source LLM benchmark. However, it struggles to differentiate between the various MISC codes, often leading to the generation of unintended outputs that deviate from the given prompt. We set the hyper-parameter temperature as 0 to control the randomness of generation and ensure reproducibility.

4.4 Prompt template design

We elaborate a comprehensive prompt template by the following steps: (1) We craft an initial prompt template, containing task instructions, MISC guidelines, illustrative examples, and historical conversations; (2) We conduct meticulous manual verification and refinement until the resulting output codes meet our predefined expectations. An example of the prompt template is detailed in Appendix A. Notably, there are two integral components within the prompts:

MISC coding manual: We introduce the definition of a list of role-specific codes with their names and comprehensive descriptions (see Table 2). The role of the current speaker is either therapist or client. The descriptions are carefully crafted and condensed into a scheme handbook (Miller et al., 2002)* by experts specializing in MI in psychotherapy.

MISC coding examples: We offer each code two examples of client-therapist utterance pairs. These examples are selected from the MISC scheme handbook *.

5. Outcomes

5.1 Overall performance (RQ1)

To address RQ1, we conduct experiments on AnnoMI and BiMISC datasets respectively, and evaluate the performance on coarse-grained codes and fine-grained codes.

Table 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. On the BiMISC dataset, GPT-4 + mapping (predicting fine-grained codes and mapping them into coarse-grained codes) achieves substantial improvements or shows comparable results when compared to GPT-4 (predicting coarse-grained codes directly). Specifically, F1 scores on ST, QS, RF, TI increase 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 hold significant value for therapists in delivering effective MI treatment within the realm of psychotherapy. The only exception is on NT. Analyzing Figure 4, we observe that the fine-grained classification (b) has a higher false negative rate and a lower true positive rate for the NT code compared to coarse-grained classification (a). This is due to the fact that the definition of NT is not as

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

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

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

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

Dataset	Model	Marco 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 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.

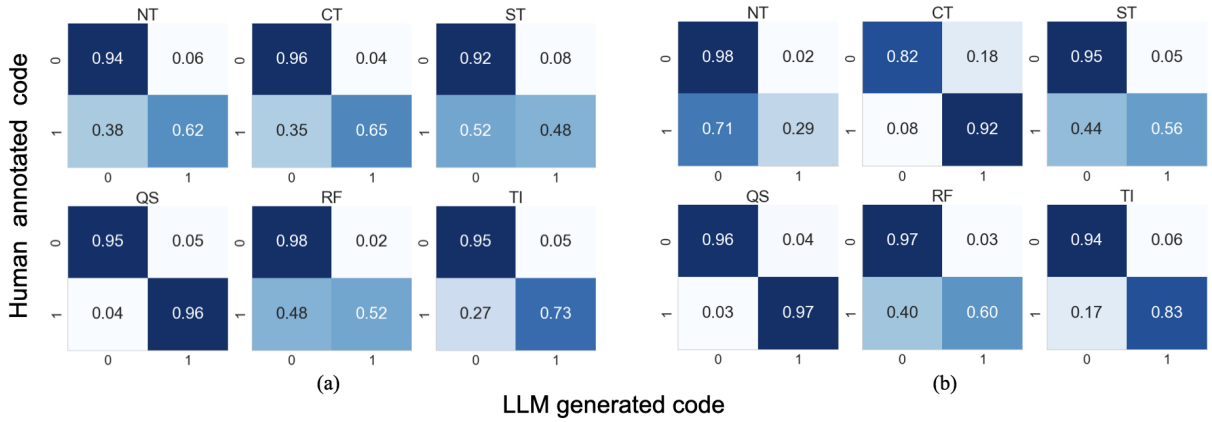

Figure 4: Accuracy of multi-code classification on the BiMISC dataset by GPT-4, using coarse-grained codes (a) and fine-grained codes (b).

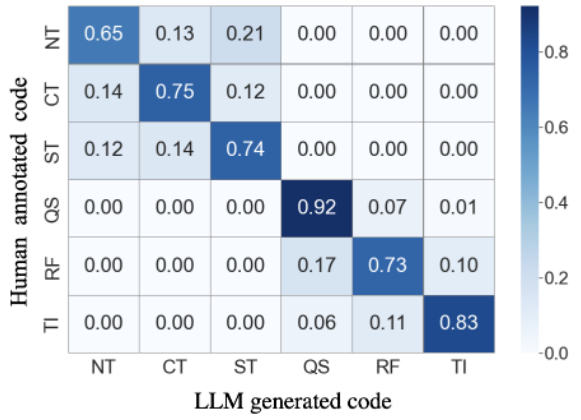

Figure 5: Accuracy of single-code classification on the AnnoMI dataset as performed by GPT-4.

clear-cut as CT and ST and it is frequently disregarded by LLMs in multi-code scenarios.

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

Third, multi-code classification (See Figure 4) is generally more challenging than single-code classification (See Figure 5). The true positive prediction accuracy for multi-code classification is typically quite high, but it is important to note that false negatives can be relatively high in certain scenarios. For example, GPT-4+mapping achives 0.71, 0.44, and 0.40 for NT, ST and RF.

5.2 Elicitation of MISC expertise (RQ2)

To address RQ2, we conduct an ablation study to access how two key factors (i.e., MISC manual and examples) influence the elicitation of MISC expertise from LLM, as shown in Table 6.

First, the choice of prompt substantially influences the LLM's performance. Specifically, in GPT-4 when using the MISC manual, the F1 scores increase from 0.58 to 0.73, marking a 15% improvement. This confirms that MISC manual can elicit

¶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.

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 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.

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 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.

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 increases by 7%, while Flan-T5 sees a 1% increase when two examples are used without the MISC manual.
Third, from a model performance standpoint, GPT-4 leads the pack, followed by Flan-T5, with GPT-3.5 comes the last. Consistent results across different prompt setups with these three models affirm the generalizability of LLMs for MISC classification and support our findings.

5.3 Multi-lingual analysis (RQ3)

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

The results, as displayed in Table 7, show 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 the GPT-4's robust multi-lingual capabilities. The multi-lingual consistency of GPT-4 in MI coding, irrespective of the language, indicates 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 provides affirmative evidence. Consistent multi-lingual performance in MISC classification paves the way for broader applications in multi-lingual psychotherapy contexts.

6. Conclusion and Future work

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

Acknowledgement

We would like to thank all research members of the TIMELY project for their valuable insights and input. This study is funded by the European Commission in the Horizon H2020 scheme, awarded to the TIMELY project (Grant agreement ID: 101017424). We also thank our anonymous reviewers for their comments.

Limitations

While our work introduces 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.

Reproducibility

To promote the development of MISC coding research and facilitate the reproducibility of the research, we release the work of BiMISC at the repository: https://github.com/XIN-von-SUN/Eliciting-MISC-in-Psychotherapy-with-LLMs.git.

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 implement rigorous data anonymization procedures. All identifiable information, including names, addresses, and any specific personal details, are meticulously removed or replaced with pseudonyms to ensure confidentiality and anonymity.

Expert Annotation

To maintain the integrity and quality of the data, annotation is 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 acknowledge and carefully consider 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. Future researchers who wish to utilize the dataset are expected to adhere to these ethical standards and guidelines, ensuring that the data is used responsibly and ethically for the advancement of knowledge in the field.

References

Dion Alperstein and Louise Sharpe. 2016. The efficacy of motivational interviewing in adults with chronic pain: A meta-analysis and systematic review. *The Journal of Pain*, 17(4):393–403.

David Atkins, Mark Steyvers, Zac Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation science : IS*, 9:49.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dogan Can, David Atkins, and Shrikanth Narayanan. 2015. A dialog act tagging approach to behavioral coding: a case study of addiction counseling conversations. pages 339–343.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. pages 5599–5611.

Rob Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. Llm-assisted content analysis: Using large language models to support deductive coding.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023a. Syllogistic reasoning for legal judgment analysis. In *Empirical Methods in Natural Language Processing (EMNLP'23)*.

Wentao Deng, Jiahuan Pei, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023b. Intent-calibrated self-training for answer selection in open-domain dialogues. *Transactions of the Association for Computational Linguistics*, 11:1232–1249.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. 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)*, pages 3816–3830, Online. Association for Computational Linguistics.

Google. 2023. Google translate.

Zac Imel, Derek Caperton, Michael Tanana, and David Atkins. 2017. Technology-enhanced human interaction in psychotherapy. *Journal of Counseling Psychology*, 64.

Cas P. D. R. Schaap Jannet M. de Jonge, Gerard M. Schippers. 2005. The motivational interviewing skill code: Reliability and a critical appraisal. *Cambridge University Press*.

Florian Klonek, Vicenç Quera, and Simone Kauffeld. 2015. Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior*, 44.

Diane Korngiebel and Sean Mooney. 2021. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. *npj Digital Medicine*, 4.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35.

Renata Martins and Daniel McNeil. 2009. Review of motivational interviewing in promoting health behaviors. *Clinical psychology review*, 29:283–93.

William Miller and Stephen Rollnick. 2002. Motivational interviewing: Preparing people for change (2nd ed.). *Journal For Healthcare Quality*, 25:46.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2002. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. 2016. The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity. *J Subst Abuse Treat*, 65:36–42.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Juri Opitz and Sebastian Burst. 2021. Macro f1 and macro f1.

Stefano De Paoli. 2023. 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.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. 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*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu.

2020. Exploring the limits of transfer learning with a unified text-to-text transformer. 21(1).

Tim Rietz and Alexander Maedche. 2021. Cody: An ai-based system to semi-automate coding for qualitative research.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. volume Vol. 4304, pages 1015–1021.

Alexander Street. 2023. Alexander street.

Robert H. Tai, Lillian R. Bentley, Xin Xia, Jason M. Sitt, Sarah C. Fankhauser, Ana M. Chicas-Mosier, and Barnas M. Monteith. 2023. Use of large language models to aid analysis of textual data. *bioRxiv*.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, 65.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. volume 2021, pages 110–115.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Bo Xiao, Dogan Can, James Gibson, Zac Imel, David Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. pages 908–912.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. 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*, IUI '23 Companion, pages 75–78, New York, NY, USA. Association for Computing Machinery.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, and Maarten de Rijke. 2022. Remedi: Resources for multi-domain, multi-service, medical dialogues. International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22).

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. *27th International Conference on Intelligent User Interfaces*.

Wu Zixiu, Balloccu Simone, Kumar Vivek, Helaoui Rim, Reiter Ehud, Reforgiato Recupero Diego, and Riboni Daniele. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Appendix A: An example of the prompt template 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 1: Prompt template for MISC classification on the therapist's utterance.