Eyes Can't Always Tell: Fusing Eye Tracking and User Priors for User Modeling under AI Advice Conditions

Xin Sun
National Institute of Informatics
Tokyo, Japan
xin.von.sun@gmail.com

Shu Wei
Yale School of Medicine
New Haven, USA
s.wei@yale.edu

Ting Pan
University of Amsterdam
Amsterdam, Netherlands
t.pan@uva.nl

Yajing Wang
University of Amsterdam
Amsterdam, Netherlands
y.wang16@uva.nl

Jos A. Bosch
University of Amsterdam
Amsterdam, Netherlands
j.a.bosch@uva.nl

Isao Echizen
National Institute of Informatics
Tokyo, Japan
iechizen@nii.ac.jp

Abdallah El Ali
Centrum Wiskunde & Informatica
Amsterdam, Netherlands
Utrecht University
Utrecht, Netherlands
Abdallah.El.Ali@cwi.nl

Saku Sugawara
National Institute of Informatics
Tokyo, Japan
saku@nii.ac.jp

Abstract

Modeling users' cognitive states (e.g., cognitive load and decision confidence) is essential for building adaptive AI in high-stakes decision-making. While eye tracking provides non-invasive behavioral signals correlated with cognitive effort, prior work has not systematically examined how AI assistance contexts, specifically varying advice reliability and user heterogeneity, can alter the mapping between gaze signals and cognitive states. We conducted a within-subject lab eye-tracking study (N=54) on factual verification tasks under three conditions: No-AI, Correct-AI advice, and Incorrect-AI advice. We analyze condition-dependent changes in self-reports and eye-tracking patterns and evaluate the robustness of eye-tracking-based user modeling. Results show that AI advice increases decision confidence compared to No-AI, while Correct-AI is associated with lower perceived cognitive load and more efficient gaze behavior. Crucially, predictive modeling is context-sensitive: the relationship between eye-tracking signals and cognitive states shifts across AI conditions. Finally, fusing eye-tracking features with user priors (demographics, AI literacy/experience, and propensity to trust technology) improves cross-participant generalization. These findings support condition-aware and personalized user modeling for cognitively aligned adaptive AI systems.

ACM Reference Format:
Xin Sun, Shu Wei, Ting Pan, Yajing Wang, Jos A. Bosch, Isao Echizen, Abdallah El Ali, and Saku Sugawara. 2026. Eyes Can't Always Tell: Fusing Eye Tracking and User Priors for User Modeling under AI Advice Conditions. In *34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26), June 08–11, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3774935.3806190

1 Introduction

As AI becomes increasingly embedded in high-stakes human decision making [16], understanding users' moment-to-moment states, such as perceived cognitive load and decision confidence, is critical for building human-AI aligned interactive systems. In practice, the same user may respond differently [23], in attention allocation, confidence, and decisions, depending on *how* AI is introduced and *whether* the AI advice is correct [45]. Prior work in human-AI interaction shows that perceived trustworthiness [51, 59], confidence in decision-making [52, 59], and AI reliability and transparency [50, 55] influence how people engage with information, which can affect people's cognitive and attentional states. These findings motivate user modeling approaches that sense and adapt to users' cognitive states during AI-assisted decision-making.

A promising signal source for such user modeling is eye tracking: gaze behavior and pupil-related measures have been used as behavioral correlates of cognitive and attentional effort in reading, reasoning and decision-making [23, 48, 56, 61]. However, it remains unclear whether these signals provide a *robust* basis for cognitive-state modeling when the context is influenced by AI conditions. Users may allocate cognitive attention differently, adopt different strategies, and experience different levels of confidence and effort depending on whether AI advice is present and whether it is correct or not (cf., hallucinations [28, 57] or misinformation [58]). Most prior work analyzes gaze or physiological signals without explicitly accounting for the AI condition in cognitive state modeling [44, 61]. Moreover, AI-assisted decision-making exhibits substantial *individual differences* [51]. User factors such as demographics, AI literacy and experience, and propensity to trust technology can influence reliance on AI and subjective judgments [50, 51]. Yet prior work of eye-tracking-based user modeling rarely tests whether incorporating such user priors improves modeling in AI assistance, leaving open questions about cross-participant generalization and condition sensitivity of user modeling using eye-tracking signals.

To address these gaps, we investigate how eye-tracking signals relate to self-reported cognitive load, decision confidence, and decision performance under varying AI advice conditions, and whether user priors improve generalization to unseen users. We conduct a controlled lab study using factual verification tasks with $N = 54$ participants under three within-subject conditions: (1) **No-AI**, (2) **Correct-AI** advice, and (3) **Incorrect-AI** advice. For each factual claim, we created matched Correct and Incorrect AI advice, where Incorrect-AI *flipped the recommendation relative to the ground-truth* while keeping length and format comparable to Correct-AI advice,

*Corresponding authors: Abdallah El Ali; Saku Sugawara.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
UMAP '26, Gothenburg, Sweden
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2311-7/26/06
https://doi.org/10.1145/3774935.3806190

and we verified intended correctness manipulation via internal validation. For each trial, we collect self-reports and simultaneously record eye-tracking signals (i.e., gaze behaviors and pupil measures). We evaluate predictive models under leave-one-subject-out cross-validation (LOSOCV) [25] to assess cross-participant generalization and report condition-stratified performance.

Our investigation is guided by three research questions:

RQ1: *How do AI presence and correctness affect users' cognitive load, decision confidence and accuracy, as well as eye-tracking patterns?*

RQ2: *Can models reliably predict self-reports (cognitive load and confidence) and decision accuracy from eye-tracking signals, and how does performance vary across AI conditions?*

RQ3: *Do user priors (demographics, AI literacy and experience, and propensity to trust technology) improve cross-participant generalization of user modeling when fused with eye-tracking features?*

Our results reveal three key findings. First, AI advice influences both explicit self-reports and implicit eye-tracking signals: compared to No-AI, AI advice increases decision confidence, and *Correct-AI* is associated with lower perceived cognitive load and more efficient visual processing patterns. Second, user modeling from eye tracking is *condition-sensitive*: models trained on pooled data do not consistently generalize across conditions, and the best-performing features/models differ across conditions and predicted targets. Third, fusing user priors (demographics, AI literacy and experience, and propensity to trust technology) with eye-tracking features improves cross-participant generalization.

This work makes three contributions: (1) We characterize condition-sensitive effects in self-reported cognitive states and eye-tracking patterns under varying AI assistance. (2) We systematically compare and evaluate pooled and condition-specific modeling strategies under AI assistance and correctness manipulation. (3) We show that fusing user priors with eye-tracking signals improves generalization, offering practical implications for cognitively-aligned user modeling and adaptive AI-assisted decision-making systems that account for both AI conditions and individual differences.

2 Related Work

2.1 AI-assisted decision-making

AI is increasingly used to support human decision-making, and carefully designed assistance can improve decision performance and trust in some settings [46, 59]. At the same time, AI support can induce reliance and errors: automation bias and related overreliance effects lead users to accept incorrect advice or miss AI failures, especially when reliability cues are unclear [38, 50]. A growing body of HAI work shows that users' judgments and reliance are affected by how AI *communicates* its trustworthiness, e.g., certainty indicators, transparency, and explanations, and that these cues can have context-dependent effects on perceived trust and downstream decisions [30, 43, 55, 59]. Recent studies further highlight that AI-provided confidence can shift users' *self-confidence* and its calibration, potentially persisting beyond the interaction [29], and that explanation quality or even misleading/imperfect explanations can meaningfully alter users' acceptance and trust [26, 27, 34].

2.2 Gaze signal for cognitive state inference

Eye-tracking provides a non-invasive window into cognitive processing and decision-making [56]. A large body of psychology and HCI research has established that gaze behaviors such as longer fixations, and physiological signals such as pupil dilation can reflect processing difficulty, mental effort, and cognitive load [23, 32, 61]. Besides, prior work has operationalized eye-tracking features (e.g., fixation/saccade statistics, pupil dynamics) as inputs for machine learning models to infer users' perceived trust [3], sources [2, 44, 60], cognitive load, and biases [7, 23, 32, 36] in response to AI-generated content. These results suggest that gaze signals can support low-burden sensing for user modeling when the mapping between attention/physiology and subjective state is relatively stable.

However, applying gaze-based inference to AI-assisted decision-making raises a key challenge of *signal attribution*. In such tasks, an observed change in gaze or pupil dilation may reflect multiple overlapping sources: inherent evidence difficulty, additional effort spent interpreting AI advice or transparency cues, or cognitive conflict when AI suggestions are misleading [7]. Meanwhile, AI assistance can also shift users' decision strategies and attention allocation as a function of perceived reliability, trustworthiness and reliance [50, 51, 59]. This implies that the mapping from implicit gaze signals to subjective states may *change with AI context* rather than remain fixed, making it unclear whether models trained in one setting will generalize across different reliability conditions or across users. Our work directly addresses this gap by testing gaze-based inference of cognitive load and decision confidence under controlled No-AI, Correct-AI, and Incorrect-AI contexts.

2.3 Personalized and multimodal user modeling

UMAP research emphasizes modeling users' states and individual differences to support personalization and adaptive interaction. Personalized user modeling aims to capture individual differences (e.g., skills, attitudes, trust) so that systems can adapt to *who* the user is, not only *what* they do in a single interaction. In AI-assisted decision-making, such personalization is especially important because users vary in AI literacy and experience [9] and in their propensity to trust automation/technology [20, 22], which can shape reliance, confidence, and susceptibility to overreliance. Recent HAI work further argues that responsible AI experiences should account for heterogeneity in users' beliefs and interpretations of system cues [30, 31].

A complementary line of work uses data to infer user states by combining behavioral and physiological signals and learning predictive models [3–5, 23, 32, 44]. However, sensing-based models emphasize within-subjects performance, while deployable settings require generalization to unseen users (a cold-start problem). This motivates us to fuse *dynamic interaction signals* (e.g., gaze) with *stable user priors* (e.g., demographics, AI literacy, trust propensity) to improve cross-participant robustness and calibration.

3 Methods

We conducted experiments with the following three steps (Fig. 1).

3.1 Empirical study with eye-tracking

3.1.1 Study design and conditions. As shown in Fig. 1, we used a within-subjects design in which each participant completed 12

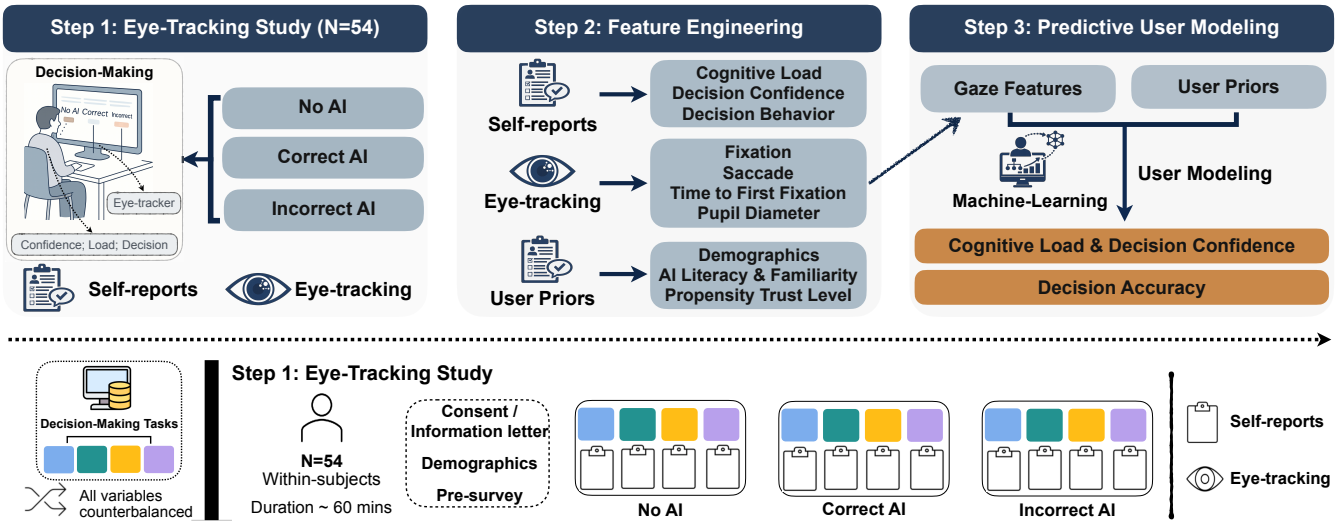


Figure 1: Study overview and procedure. ***(Top):*** **Three-step workflow.** ***Step 1*** **collects eye-tracking signals and self-reports during factual verification under three within-subject AI conditions.** ***Step 2*** **extracts trial-level eye-tracking signals and outcomes (cognitive load, decision confidence, and accuracy) together with participant-level user priors (e.g., demographics, AI literacy, propensity to trust).** ***Step 3*** **trains machine learning models to predict users' cognitive load, decision confidence, and accuracy from eye-tracking features alone or fused with user priors, across AI conditions.** ***(Bottom):*** **Study procedure in Step 1: consent and pre-survey, followed by counterbalanced trials spanning three AI conditions with concurrent eye-tracking and self-reports.**

factual verification trials as the decision-making. In each trial, participants read a claim and its supporting evidence (adapted from dataset StrategyQA [13]) and then made a binary *True/False* judgment. To manipulate the decision context, each trial was assigned to one of three AI assistance conditions: (1) *No-AI*: no AI advice was shown; (2) *Correct-AI*: correct AI-generated advice shown; (3) *Incorrect-AI*: an incorrect AI suggestion was shown. Trials were evenly distributed across three AI conditions (No-AI / Correct-AI / Incorrect-AI; 4 each). Each participant experienced all conditions in a counterbalanced order to mitigate sequence effects.

3.1.2 Participants. A priori G*Power [12] analysis suggested 32 participants to detect a medium effect (Cohen's $d = 0.25$) with 80% power. We recruited *54* participants via the institute's subject pool. All participants received compensation for a 60-minute in-lab session. The study was approved by the institute's ethical committee. Participant demographics are reported in Table 1.

Demog.	Categ.	Participants (%)
Gender		
	Female	48 (88.9%)
	Male	6 (11.1%)
Age		
	18-24	36 (66.7%)
	25-34	16 (29.6%)
	35-44	2 (3.7%)
Education		
	High school	8 (14.8%)
	Bachelor's degree	28 (51.9%)
	Master's degree	17 (31.5%)
	Doctorate	1 (1.9%)
AI Familiarity		
	Never	0 (0.0%)
	Slightly	5 (9.3%)
	Moderately	20 (37.0%)
	Very	21 (38.9%)
	Extremely	8 (14.8%)

Table 1: Characteristics of the participants in Study of Step 1.

3.1.3 Materials: stimuli and apparatus. We developed a custom web-based interface to present stimuli and collect participants' responses. Each task consisted of a binary fact-checking claim (True/False) with supporting context. The interface displayed the LLM-generated answer. After each trial, participants provided their self-reported decisions and ratings. The factual claims were adapted from the public StrategyQA dataset [13]. For each claim, we prepared two AI advice variants matched in length and format: (i) Correct-AI advice aligned with the ground-truth, and (ii) Incorrect-AI advice whose recommendation was intentionally flipped to be incorrect while keeping surface-level style comparable. We conducted an internal check to ensure the manipulation of advice. Examples of the claim and AI advice are added in the supplementary material.

Fig 2 shows that experiment was run on a Windows PC with a 27" monitor (1920 x 1080, 100 Hz). A Tobii Pro Fusion eye tracker (60 Hz) recorded eye-tracking signals [47]. Stimuli and questionnaires were presented through PsychoPy [37], which synchronized experimental tasks with recording to ensure precise alignment.

3.1.4 Measure. **User priors.** Prior to the study, participants completed a short questionnaire capturing demographics and user priors related to human-AI interaction (Table 1), including prior experience with AI, AI literacy [9] and propensity to trust technology (PPT) [22] (used as participant-level features in modeling).

Self-reports. After each task, participants provided the following responses: (1) **Decision:** Participants made a binary decision on the given factual claim (True or False). (2) **Decision confidence:** Participants rated how confident they were in their decision on a 7-point Likert scale (from "Not confident at all" to "extremely confident") [52]. (3) **Cognitive load:** Participants reported perceived cognitive load after each trial using 7-point Likert scales (1=low, 7=high) adapted from [35] (Cronbach's alpha=.79). An additional

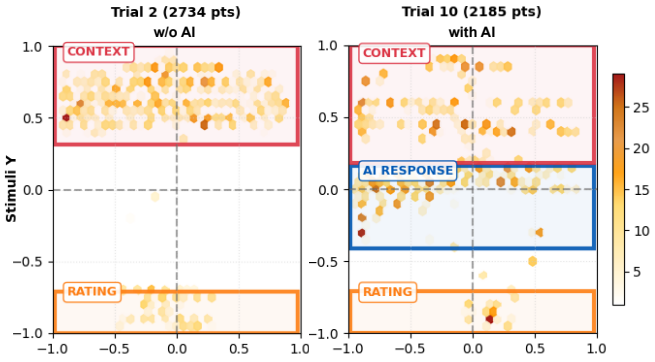


Figure 2: Heatmaps of gaze per trial from conditions with or without AI. Higher density indicates greater visual attention.

Manipulation check was used after each AI trial to check whether participants perceived the AI response as correct or not.

Eye-tracking signals. We recorded gaze behaviors and physiological pupil measures throughout trials. We derived standard eye-tracking features, including: *fixation count and duration [49], saccade count and length [53], time to first fixation (TTFF), and pupil diameter [23, 32].* Areas of interest (AOIs) were defined as Fig. 2 (i.e., claim and evidence text as AOI-Context, AI advice panel when present as AOI-Advice, and the rating panel as AOI-Rating).

3.1.5 Procedure. As shown in Fig. 1, participants provided consent and completed a short pre-survey. They then sat in front of an eye-tracking setup and completed a standard eye-tracker calibration routine [21]. The study began with a brief tutorial trial to familiarize participants with the study. Each participant completed 12 trials. In each trial, participants read claim and supporting evidence, viewed AI advice when present, and made a binary decision. They then reported their decision confidence and perceived cognitive load.

3.1.6 Data analysis. We analyzed the data to address RQ1.

Preprocessing of eye-tracking data. We applied standard preprocessing [21]. Gaze patterns and pupil diameter were calculated by the I-VT signal filter [40]. We excluded trials that did not meet predefined quality criteria (gaze validity < 80%).

Condition-aware effects. We analyzed self-reports (cognitive load, confidence), decision accuracy by repeated measures ANOVA analysis [15] with FDR-correction [19] across AI conditions to answer RQ1. Eye-tracking data was analyzed by MixedLM [1] across three AOIs. Both self-reports and eye-tracking were further analyzed by post-hoc pairwise t-test [39].

3.2 Feature engineering for predictive modeling

For each trial i, we extracted trial-level gaze and pupil features (see Sec. 3.1.4). Each trial is represented by an eye-tracking feature vector $\mathbf{g}_i \in \mathbb{R}^d$ with an associated AI condition label $c_i \in \{\textit{No-AI}, \textit{Correct-AI}, \textit{Incorrect-AI}\}$. Specifically, we computed standard eye-tracking metrics per AOI, including fixation count and duration, saccade count and length, time-to-first-fixation (TTFF), and physiological feature pupil diameter per AOI. Trial-level outcomes include participants' decision accuracy accuracy_i and self-reported confidence confidence_i and cognitive load load_i. In addition, each participant u provides user-level priors $\mathbf{p}_u$ (see Sec. 3.1.4), which are constant across that participant's trials. We use these priors either alone or fused with eye-tracking features for cross-participant modeling. Thus, our dataset consists of trial-level features:

$$\mathcal{D} = \left\{\left(\mathbf{c}_i,\ \mathbf{g}_i,\ p_u,\ \text{load}_i,\ \text{confidence}_i,\ \text{accuracy}_i\right)\right\}_{i=1}^{N},$$

3.3 Conditional and personalized user modeling

We formulate user modeling as a supervised prediction task: given eye-tracking features and user priors, we predict participants' self-reported *cognitive load*, *decision confidence* and decision accuracy.

Modeling across AI conditions (RQ2). Each trial is labeled with an AI condition $c_i \in \{\textit{No-AI}, \textit{Correct-AI}, \textit{Incorrect-AI}\}$. We evaluate three modeling strategies: **(1) Pooled (All):** one model trained on all trials across conditions. **(2) Condition-specific:** three models trained within each condition. **(3) Condition-aware mixture-of-experts (MoE):** three condition-specific experts trained separately, where inference routes each test trial to the expert corresponding to its known condition c_i. *To address **RQ2***, we trained machine learning (ML) models to learn the mapping:

$$f : \mathbf{g}_i \mapsto \left\{\left(\text{load}_i, \text{confidence}_i, \text{accuracy}_i\right)\right\}_{i=1}^{N},$$

Personalized modeling with priors (RQ3). We examined ML prediction with the value of user priors across AI conditions. We evaluate three feature settings: **(i) Eye-tracking-only**: $\mathbf{g}_i$, include both gaze behaviors (i.e., fixation counts/duration, saccade counts/length, TTFF) and physiological pupil diameter; **(ii) User Priors-only**: $\mathbf{p}_{u_i}$, include demographics, AI literacy and experience, and propensity to trust technology (Sec. 3.1.4).; **(iii) Multimodal Fusion**: $[\mathbf{g}_i + \mathbf{p}_{u_i}]$. This ablation isolates how much predictive signal comes from eye-tracking versus user priors, and whether they provide complementary information. Thus, we trained machine learning (ML) models to learn the mapping:

$$f : (\mathbf{g}_i, \mathbf{p}_{u_i}) \mapsto \left\{\left(\text{load}_i, \text{confidence}_i, \text{accuracy}_i\right)\right\}_{i=1}^{N},$$

Setup and evaluation. We binarize the self-reported *cognitive load* and *decision confidence* for classification. For each training fold, we compute the median of the predicted target on the *training participants only* and use it as the threshold: trials with values above (or equal to) the training median are labeled as *high*, and those below as *low*, to ensure label balance. We report Accuracy, F1 [11], and AUC [33] as the evaluation metrics.

4 Results

4.1 Effects of AI-assisted decision-making - *RQ1*

Measure	ANOVA	Pairwise (mean)	p	Effect
Load	.001 (.33)	No vs. Corr (3.56 vs. 3.18)	.010	.35
		No vs. Incorr (3.56 vs. 3.38)	.200	.18
		Corr vs. Incorr (3.18 vs. 3.38)	.040	.28
Confidence	<.000 (.42)	No vs. Corr (5.22 vs. 5.93)	<.001	.60
		No vs. Incorr (5.22 vs. 5.79)	.002	.45
		Corr vs. Incorr (5.93 vs. 5.79)	.220	.17
Accuracy	.500 (.12)	No vs. Corr (.78 vs. .77)	.830	.03
		No vs. Incorr (.78 vs. .82)	.270	.15
		Corr vs. Incorr (.77 vs. .82)	.080	.24

Table 2: Results from ANOVA analysis (p value with effect size) and pairwise comparisons with mean values. Measures are cognitive load, decision confidence and decision accuracy.

4.1.1 Self-reports. Table 2 reports the ANOVA and pairwise comparisons across No-AI, Correct-AI, and Incorrect-AI conditions. The manipulation check was significant ($p < .001$), indicating that participants perceived the intended advice correctness differences.

Cognitive load differed significantly by condition (ANOVA: $p = .001$, $\eta^2 = .33$). Pairwise tests showed lower load in *Correct-AI* than *No-AI* (3.18 vs. 3.56, p=.010, effect=.35) and than *Incorrect-AI* (3.18 vs. 3.38, p=.040, effect=.28). The difference between *No-AI* and *Incorrect-AI* was not significant (3.56 vs. 3.38, p=.200).

Decision confidence differed significantly by condition (ANOVA: $p < .001$, effect=.42). Confidence was higher with AI advice than without AI for both Correct-AI (p<.001, effect=.60) and Incorrect-AI (p=.002, effect=.45), with no significant difference between *Correct-AI* and *Incorrect-AI* conditions (p=.220). **Decision accuracy** also did not differ significantly by condition (ANOVA: p=.500, effect=.12).

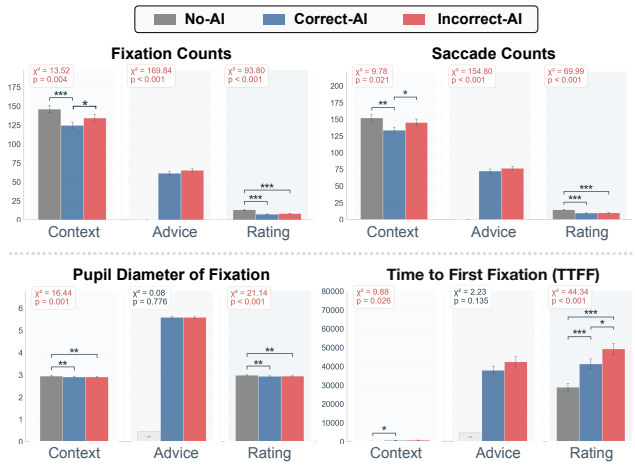


Figure 3: Gaze features (i.e., Fixation/Saccade Count, Pupil Diameter, and Time-to-First-Fixation across AOIs (Context, AI Advice, and User Rating areas). Bars show the mean. Values in upper-left (X^2, p) are results from MixedLM model (FDR-corrected), and brackets indicate significant pairwise t-tests ($*p < .05, **p < .01, ***p < .001$).

4.1.2 Gaze patterns vary across conditions. We processed the raw eye-tracking signal [21] and analyzed it to complement self-reported outcomes and test how presence and correctness of AI advice shape visual attention (Fig. 3) across conditions and AOIs (see Sec. 3.1.6).

No-AI. Participants showed more and longer fixations and saccades, as well as larger pupil diameter on the context (AOI-Context) and rating items (AOI-Rating), than in conditions with AI. Participants also oriented most quickly to rating items (shorter time to first fixation). This pattern suggests that participants relied primarily on the context for decisions with higher uncertainty, consistent with the lowest decision confidence and highest cognitive effort.

Correct-AI. With correct AI advice, participants made significantly fewer fixations and saccades on the evidence (AOI-Context) and rating items (AOI-Rating) than in No-AI condition. Compared to Incorrect-AI, they showed lower saccade activity in advice region (AOI-Advice), suggesting reduced effort when processing advice. These patterns are consistent with self-reports showing higher decision confidence and lower cognitive load in Correct-AI condition.

Incorrect-AI. Participants showed more visual processing on AOI-Context than Correct-AI condition, reflected by higher fixations and saccades. Participants also took longer to shift attention to the AOI-Rating, with the *longest* TTFF in Incorrect-AI. It indicates that Incorrect-AI is associated with increased cognitive effort to evidence when processing incorrect AI advice.

4.2 User modeling via eye-tracking alone - *RQ2*

Table 3 (Eye-tracking Only part) answers **RQ2** on whether eye-tracking signals alone can reliably predict (i) participants' self-reported *cognitive load* and *decision confidence*, (ii) *decision accuracy*, and how predictions vary across AI conditions. We compare three modeling settings (Sec.3.3): (1) *pooled* training over all AI conditions (*All*), (2) *condition-specific* training within each AI condition (No-AI, Correct-AI, Incorrect-AI), and (3) a condition-aware *MoE*.

Across predictive targets, eye-tracking signals are *most predictive of decision accuracy*. Under pooled training (*All*), all models (except LR) reach high accuracy ($\sim$ 0.79), indicating that eye-tracking carries robust information about whether a decision is correct. However, the best performance often emerges under *condition-specific* training rather than pooling: for some models, accuracy increases to low $\sim$ 0.8 range (e.g., $\sim$ 0.82 (ET) - $\sim$ 0.83 (SVM)) in certain AI conditions. This pattern suggests that performance of modeling from eye-tracking is not fully invariant across conditions.

Predicting self-reported **cognitive load** and **confidence** from eye-tracking alone is more challenging. Overall accuracies for these predictions are lower and show larger fluctuations across AI conditions than predicting decision accuracy. For cognitive load, models achieve moderate performance (roughly $\sim$ 0.66 accuracy), and the pooled training settings achieved the best performance across most models, implying that cognitive load inferred from eye-tracking signals is sensitive to contextual factors. For decision confidence, predictive performance varies more noticeably by condition: most models achieve their highest confidence prediction in Correct-AI condition, while performance can drop in No-AI or Incorrect-AI settings for several models, consistent with the idea that misleading AI advice can disrupt the stability of gaze-based confidence cues.

These results answer **RQ2** in two ways. First, eye-tracking signals alone can predict *decision accuracy* relatively reliably, whereas self-reports are moderately predictable. Second, performance varies across AI conditions, and pooled (*All*) training does not always dominate condition-specific modeling. This provides direct evidence of *conditional heterogeneity*: user modeling by eye-tracking cues in AI-assisted decision-making should explicitly account for AI conditions (e.g., whether AI advice is present and whether it is correct), rather than assuming a single eye-tracking to self-report mapping transfers uniformly across AI conditions and settings.

4.3 User modeling by multimodal fusion - *RQ3*

We next examine **RQ3** on whether fusing user priors improves predictive modeling beyond eye tracking alone. We compare three feature settings (Sec. 3.3): *eye-tracking only* (g_i), *user priors only* (p_{u_i}), and *multimodal fusion* (eye-tracking + user priors; $[g_i; p_{u_i}]$).

Fusing priors with eye tracking yields the most consistent gains for cognitive load and confidence across conditions, as shown in Table 3 & 4. Across predicted targets, multimodal fusion is competitive with

Targets	Models	Eye-tracking Only					User Priors	Multimodal Fusion				
		All	No	Corr	Incorr	MoE	All	All	No	Corr	Incorr	MoE
Cognitive Load	LR	.61/.65/.64	.61/.57/.68	.50/.51/.52	.63/.60/.62	.58/.63/.63	.65/.69/.68	.66/.71/.73	.65/.64/.70	.58/.63/.62	.70/.72/.74	.65/.71/.69
	SVM	.62/.76/.51	.57/.64/.52	.60/.75/.55	.49/.45/.48	.62/.76/.52	.68/.75/.63	.66/.76/.51	.57/.63/.52	.60/.75/.56	.49/.45/.48	.61/.76/.52
	RF	.66/.74/.68	.64/.64/.68	.58/.67/.58	.60/.60/.61	.64/.74/.62	.65/.71/.68	.71/.78/.76	.67/.67/.72	.59/.69/.62	.63/.62/.66	.65/.75/.67
	ET	.65/.74/.68	.63/.62/.68	.58/.68/.57	.55/.51/.61	.62/.72/.63	.68/.74/.68	.65/.81/.79	.69/.69/.74	.63/.72/.67	.68/.75/.72	.68/.75/.72
	AdaB	.62/.73/.63	.63/.62/.67	.60/.75/.52	.57/.51/.57	.61/.70/.60	.66/.73/.67	.65/.74/.67	.67/.68/.68	.60/.72/.52	.57/.52/.56	.63/.71/.63
	XGB	.66/.75/.69	.61/.64/.66	.57/.69/.60	.62/.59/.63	.60/.70/.60	.68/.73/.67	.69/.77/.74	.65/.66/.70	.60/.71/.60	.56/.56/.62	.63/.71/.62
	MLP	.64/.74/.62	.58/.60/.62	.54/.65/.52	.57/.60/.60	.57/.66/.59	.64/.74/.67	.73/.79/.76	.61/.59/.67	.46/.51/.47	.62/.61/.65	.62/.73/.62
Decision Confidence	LR	.62/.69/.63	.63/.66/.68	.60/.68/.58	.59/.68/.59	.58/.66/.57	.57/.64/.61	.62/.68/.66	.69/.72/.73	.58/.68/.54	.62/.71/.60	.60/.68/.60
	SVM	.65/.78/.59	.66/.70/.56	.67/.79/.59	.68/.81/.54	.65/.77/.52	.65/.78/.57	.63/.76/.60	.66/.70/.56	.68/.80/.60	.68/.81/.54	.65/.77/.52
	RF	.65/.76/.67	.67/.67/.70	.71/.81/.67	.67/.78/.65	.67/.77/.66	.59/.66/.63	.66/.77/.68	.67/.71/.73	.71/.81/.69	.67/.78/.65	.67/.77/.67
	ET	.65/.77/.66	.63/.70/.69	.71/.81/.66	.66/.77/.65	.66/.77/.65	.65/.75/.63	.67/.77/.66	.68/.71/.74	.74/.83/.70	.68/.77/.61	.67/.77/.68
	AdaB	.67/.79/.64	.59/.66/.61	.70/.82/.65	.67/.80/.52	.61/.72/.59	.64/.78/.59	.68/.78/.65	.62/.68/.65	.70/.81/.64	.66/.79/.53	.63/.74/.61
	XGB	.63/.75/.65	.61/.67/.69	.71/.81/.69	.67/.79/.62	.65/.75/.64	.65/.78/.62	.64/.75/.68	.68/.72/.71	.74/.83/.72	.65/.77/.60	.63/.73/.62
	MLP	.65/.76/.64	.64/.70/.65	.65/.77/.54	.63/.74/.57	.59/.69/.58	.62/.75/.54	.65/.75/.63	.66/.70/.71	.74/.83/.62	.66/.78/.58	.66/.76/.64
Decision Accuracy	LR	.57/.69/.59	.56/.67/.53	.52/.65/.51	.54/.68/.48	.61/.73/.54	.60/.71/.62	.64/.75/.62	.59/.71/.54	.66/.77/.57	.68/.79/.55	.67/.78/.60
	SVM	.79/.88/.49	.78/.87/.41	.78/.87/.44	.83/.91/.46	.79/.88/.47	.80/.88/.61	.79/.88/.49	.78/.87/.41	.77/.87/.41	.83/.91/.49	.79/.88/.47
	RF	.79/.88/.62	.76/.86/.47	.74/.85/.55	.82/.90/.57	.79/.88/.58	.67/.77/.64	.79/.88/.64	.76/.86/.51	.76/.86/.56	.84/.91/.57	.79/.88/.61
	ET	.79/.88/.62	.76/.86/.61	.75/.86/.50	.82/.90/.61	.79/.88/.60	.80/.88/.62	.79/.77/.66	.78/.87/.61	.74/.84/.60	.81/.89/.57	.77/.87/.61
	AdaB	.79/.88/.57	.74/.85/.54	.76/.87/.50	.82/.90/.53	.75/.85/.55	.79/.88/.60	.79/.88/.56	.76/.86/.51	.76/.87/.52	.83/.90/.53	.73/.84/.55
	XGB	.78/.88/.62	.77/.87/.58	.73/.84/.47	.82/.90/.59	.74/.85/.57	.80/.88/.63	.78/.88/.65	.76/.86/.57	.75/.86/.47	.83/.91/.57	.75/.85/.56
	MLP	.79/.88/.54	.72/.83/.51	.67/.79/.43	.79/.88/.48	.70/.82/.53	.79/.88/.58	.79/.88/.55	.73/.84/.53	.71/.82/.54	.85/.91/.58	.78/.87/.62

Table 3: Predictive modeling results under different feature sets and AI conditions. Binary classification performance (Accuracy/F1/AUC) for predicting *cognitive load, decision confidence and accuracy*. We compare three feature settings: *Eye-tracking only, User priors only*, and *Multimodal fusion* (eye tracking + user priors). For each predicted target, we report overall performance on pooled trials (*All*) and condition-stratified performance on specific *No-AI, Correct-AI*, and *Incorrect-AI* trials; *MoE* denotes a condition-aware mixture-of-experts model. Boldface indicates the best condition (All/No-AI/Correct-AI/Incorrect-AI/MoE) for each model within a given feature setting. Red highlights the best-performing values for each predicted target within that given feature setting across conditions. Model acronyms: LR (Logistic Regression); SVM (Support Vector Machine); RF (Random Forest); ET (ExtraTrees); AdaB (AdaBoost); XGB (XGBoost); MLP (Multi-Layer Perceptron).

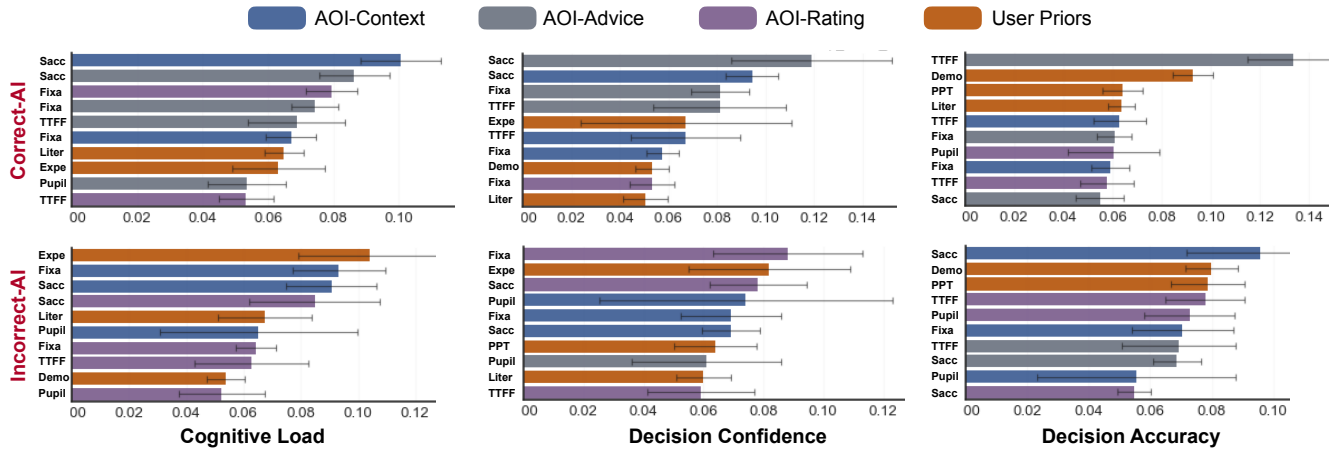


Figure 4: SHAP analysis: top 10 important features in user modeling on two AI conditions (Correct-AI vs. Incorrect-AI), to predict self-reported cognitive states (cognitive load and confidence) and decision accuracy by ExtraTree classifiers. ("Sacc" = saccades; "Fixa" = fixations; "Pupil" = Pupil Diameter; "Liter" = AI literacy; "Expe" = AI experience; "Demo" = Demographics;).

or better than priors or gaze only, with the clearest gains for predicting **cognitive load** and **decision confidence**. For example, under pooled training (*All*), fusion raises cognitive-load performance relative to eye-tracking only for several classifiers (e.g., MLP improves from .64/.74/.62 to .73/.79/.76), indicating that user priors add useful information that is not captured by eye-tracking patterns alone. A similar pattern appears for decision confidence, where fusion reaches higher performance (e.g., XGBoost). The gains of fusion are condition-dependent: improvements are typically larger when no advice is present than Correct-AI and Incorrect-AI, implying with the idea that user priors may complement how users attend to and utilize advice when misleading AI is introduced.

For decision accuracy, improvements from priors and fusion are smaller. This pattern suggests that while priors and eye-tracking signals are informative for modeling perceived cognitive states (*load* and *confidence*), predicting objective behaviors remains more challenging and may depend on factors beyond eye-tracking (e.g., AI experience, domain knowledge or item difficulty).

Model	Predicted Target	Best Feature	Best Condition	Acc/F1/AUC
LogReg	Cognitive Load	Fusion	Incorrect-AI	.70 / .72 / .74
	Decision Confidence	Fusion	No-AI	.69 / .72 / .73
	Decision Accuracy	Fusion	Incorrect-AI	.67 / .79 / .55
ExtraTrees	Cognitive Load	Fusion	Pooled (All)	.69 / .69 / .74
	Decision Confidence	Fusion	Correct-AI	.74 / .83 / .70
	Decision Accuracy	Fusion	Incorrect-AI	.82 / .90 / .61
XGBoost	Cognitive Load	Fusion	Pooled (All)	.73 / .79 / .73
	Decision Confidence	Fusion	Correct-AI	.74 / .83 / .72
	Decision Accuracy	Fusion	Incorrect-AI	.83 / .91 / .57
MLP	Cognitive Load	Fusion	Pooled (All)	.73 / .79 / .76
	Decision Confidence	Fusion	Correct-AI	.74 / .83 / .62
	Decision Accuracy	Fusion	Incorrect-AI	.85 / .91 / .58

Table 4: Summary of best-performing settings. We report the feature set and AI condition that yields the best performance.

User priors are informative, but work best when combined with gaze. User priors only models already provide competitive predictive power, suggesting stable individual differences in self-report tendencies. However, fusion is more consistently reliable across models and predicted targets than user priors alone, supporting a *complementarity* view: priors capture between-participant tendencies, while gaze captures within-trial behavioral signals.

Where fusion helps most depends on the AI condition, revealing condition sensitivity. Best-performing settings concentrate on specific AI conditions rather than uniformly under pooled training. The summary of best settings (Table. 3 & 4) shows a clear condition-dependent structure: (i) **decision confidence** often peaks under *Correct-AI* with fusion (e.g., several models achieve $\sim$ .74 accuracy with strong F1/AUC), while (ii) **decision accuracy** consistently peaks under *Incorrect-AI* with fusion (e.g., $\sim$ .82–.85 accuracy across multiple models). This indicates that priors are particularly helpful when the mapping from gaze to outcomes shifts with AI assistance, for instance, misleading AI can change strategies and uncertainty, making gaze alone less transferable across participants.

Overall, these findings address RQ3 that incorporating user priors improves user modeling when fused with eye-tracking, with gains that are *condition-sensitive*.

4.4 Features importance across AI conditions

Fig. 4 shows feature importance by SHAP [6] analysis for ExtraTrees models trained under *Correct-AI* and *Incorrect-AI* conditions.

For **cognitive load**, *Correct-AI* is dominated by eye-tracking-derived features (e.g., fixation/saccade-related metrics and pupil cues), suggesting that load is primarily reflected in visual processing dynamics when advice is reliable. However, under *Incorrect-AI*, **user priors** (notably AI experience) rise sharply and can become top-ranked, implying that stable individual differences help explain effort variation when users face misleading assistance.

For **decision confidence**, features related to attention on **AI advice** and **rating** regions consistently appear among the top contributors, consistent with confidence being shaped by how users evaluate AI suggestions. Importantly, priors become more prominent under *Incorrect-AI*, indicating that confidence under misleading advice are more strongly moderated by individual predispositions.

For **decision accuracy**, *Correct-AI* shows a strong contribution of AOI-Advice feature, whereas *Incorrect-AI* exhibits higher rankings on AOI-Context. High reliance on **user priors** (demographics, literacy, trust propensity) supports the interpretation that priors act as an interpretable stabilizer for cross-participant generalization.

5 Discussion

5.1 Condition-sensitive user modeling

A key takeaway from our study is that user state inference from eye-tracking is *condition-sensitive*. In our experiment, AI assistance systematically altered users' self-reported states (Table 2): relative to No-AI, both AI conditions increased decision confidence, and Correct-AI reduced perceived cognitive load. These shifts were accompanied by changes in eye-tracking patterns (RQ1) and translated into heterogeneous predictive performance across conditions (RQ2), indicating that the same sensing modality supports different levels of inference reliability depending on how AI is involved.

This condition sensitivity is also reflected in modeling outcomes. Training on pooled trials from all conditions did not consistently outperform condition-specific modeling, and the strongest results for confidence and cognitive load often emerged from specific contexts rather than the pooled set (Table 3). A plausible mechanism is that AI involvement changes users' decision strategy and attention allocation [44, 54]: when AI suggestions are present, users may skim evidence differently, rely more on the advice panel, or experience lower effort during rating; when suggestions are misleading, users may experience conflict and adopt different verification patterns (Fig. 3). Such strategy shifts align with prior human-AI interaction findings that perceived reliability, trust, and transparency affect how users engage with AI outputs and when they rely on suggestions [17, 34, 41, 50, 51, 55]. Our findings show that these condition-induced shifts are not only reflected in cognitive outcomes (e.g., confidence/load) but are also observable in implicit gaze behaviors and affect the robustness of eye-tracking-based user models.

Importantly, condition sensitivity does not mean modeling by eye-tracking is ineffective; rather, it implies that *AI conditions should be treated as a feature in user modeling*. Our condition-stratified testing demonstrates that even with the same eye-tracking features, model performance varies across AI conditions and differs by predicted targets (Table 3). This finding echoes prior work using eye tracking to infer cognitive effort in stable tasks [23, 32, 36, 60, 61] by highlighting a critical boundary condition for deployable user models: when the interaction context itself shifts due to AI assistance, signal attribution and learned mappings can shift as well.

Practical implication. For deployable user models, evaluation and deployment should be *condition-aware*. At minimum, models should report condition-stratified performance (rather than only overall accuracy) and incorporate observable AI-condition features (e.g., assistance mode, presence/format of suggestions or explanations, correctness) to avoid averaging over heterogeneous behaviors. Where condition heterogeneity is strong, simple condition-aware architectures (e.g., condition-specific heads or expert routing) can provide a robust alternative to a single pooled estimator.

Beyond improving predictive user modeling, our results motivate a design for *cognitively-aligned* AI assistance: systems should jointly model (i) moment-to-moment attention and cognitive states from behavioral signals, (ii) AI conditions (e.g., assistance mode and reliability cues), and (iii) stable individual differences (e.g., AI literacy) to decide *when*, *how*, and *for whom* to adapt assistance. Concretely, this suggests a closed-loop pipeline in which eye-tracking provides

continuous state estimates, condition-aware modeling handles condition shifts (e.g., via condition features or expert routing), and user priors provide personalized capability to the systems.

These design recommendations are consistent with prior HCI practices that reliability and trust to AI influence overreliance, engagement and user behaviors [50, 51, 55] and with sensing-based HCI work cautioning that behavioral and physiological signals require careful attribution under changing task demands and real-time inference for adaptive interaction [10, 14, 24, 44].

5.2 User priors for cold-start generalization in user modeling from eye-tracking

Our results further show that incorporating stable user priors, demographics, AI literacy, and propensity trust in technology, improves the *cross-participant* generalization of eye-tracking-based user modeling (RQ3). Under participant-level evaluation, multimodal fusion (eye-tracking + priors) consistently outperforms eye-tracking-only models (Table 3), with the clearest gains in accuracy/F1 and threshold-independent discrimination (AUC), even when accuracy changes are smaller. This pattern suggests that user priors help *calibrate and rank* users' subjective states rather than merely shifting decisions across a fixed classification threshold.

These gains are particularly meaningful given the gap between trial-level and participant-level evaluation. Trial-level k-fold cross-validation is often optimistic because models can leverage stable, person-specific baselines in eye-tracking (e.g., reading speed, fixation style, pupil dynamics) when the same individual appears in both train and test splits. Our findings indicate that user priors provide an interpretable anchor for the *cold-start* setting [42] by reducing cross-user ambiguity, i.e., they help explain why the *same* eye-tracking pattern may correspond to different reported confidence or cognitive load across individuals. This is consistent with prior human-AI interaction research showing that individual differences such as AI literacy and trust level influence how people interpret and rely on AI systems or advice [50, 51], and extends that line of work by demonstrating a concrete modeling benefit when these traits are used as priors for user cognitive state estimation.

Importantly, the improvements are not uniform across conditions. Condition-stratified tests show that fusion benefits concentrate in AI conditions with stronger heterogeneity (Table 3), where eyetracking-states mappings are less stable and pooled modeling is more likely to average over divergent strategies. Thus, user priors function as a stabilizer that compensates for condition-induced variability by encoding stable tendencies (e.g., baseline experience and literacy level) that modulate how eye-tracking relates to subjective effort and certainty. This finding also helps interpret the earlier observation that "pooled" (all) training does not always help (RQ2).

Practical implication. Our results suggest treating user priors as a principled component of cold-start state inference rather than an optional add-on. In practice, systems and user models can obtain lightweight priors via short onboarding questionnaires or inferred proxies (e.g., interaction history) and fuse them with behavioral sensing to improve generalization for new users. This recommendation is grounded in prior HAI evidence that trust and literacy affect reliance and overreliance behaviors and aligns with user modeling principles that stable user traits can complement short-term behavioral signals to support personalized user modeling.

6 Limitations and Future Work

We acknowledge several limitations of this work as follows.

First, self-reports as labels. We model self-reported cognitive load and confidence, which are subjective and may be noisy due to individual differences. We used a brief load scale instead of NASA-TLX [18] to reduce fatigue. *Future work:* We should triangulate self-reports with complementary indicators (e.g., response time) and improve reliability via repeated ratings. We should caution against treating eye-tracking as a causal proxy for cognitive states [8].

Second, task specificity. We study StrategyQA-style factual verification, which offers control but may not represent broader AI-assisted decision-making. Our main conclusions about context shifts and the value of user priors require replication in other tasks and stakes. *Future work:* We will replicate the study across decision types and test whether the same condition shifts and personalization gains hold under varied cognitive demands and stakes.

Third, fixed AI advice format. AI advice is a brief suggestion and incorrect advice is created by flipping the ground-truth. This does not cover richer AI assistance formats (e.g., explanations). *Future work:* We will vary assistance style and introduce realistic errors (partial correctness, misleading evidence, hallucinations [57]).

Lastly, sample size and participant diversity. Our participant pool may limit the power of generalized user modeling. Nonetheless, our primary findings rely on within-subject comparisons, which support the robustness of the observed condition sensitivity and fusion gains. *Future work:* We will scale to larger and more diverse samples to test stability across demographic groups.

7 Conclusion

We investigated user modeling on cognitive states from eye-tracking signals and user priors in AI-assisted decision-making across three AI conditions (No-AI, Correct-AI, Incorrect-AI). The findings show that AI advice affects both self-reported cognitive states, decision-making and eye-tracking patterns. We further reveal that user modeling from eye-tracking alone is condition-sensitive, and fusing eye-tracking with user priors (demographics, AI literacy/experience, and propensity trust) can improve cross-participant generalization for user modeling. Overall, our work suggests that cognitive-state user models should incorporate both AI conditions and individual differences to enable cognitively-aligned adaptive AI systems.

GenAI Declaration

We used AI tools (i.e., GPT-5.2) only for language editing to improve clarity and conciseness. All study design, analysis, literature review, and writing were conducted and verified by the authors.

Acknowledgments

This work was supported by the JST CREST Grant (JPMJCR2562), JST K Program Grant (JPMJKP24C2), and JST FOREST Grant (JPMJFR232R) in Japan.

References

[1] 1988. Newton Raphson and EM algorithms for linear mixed effects models for repeated measures data. *J. Amer. Statist. Assoc.* 83, 404 (1988), 1014–1022.

[2] Yasmeen Abdrabou, Elisaveta Karypidou, Florian Alt, and Mariam Hassib. 2023. Investigating User Behavior Towards Fake News on Social Media Using Gaze and Mouse Movements. https://doi.org/10.14722/usec.2023.232041

[3] Muneeb Ahmad and Abdullah Alzahrani. 2023. Crucial Clues: Investigating Psychophysiological Behaviors for Measuring Trust in Human-Robot Interaction. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 135–143. https://doi.org/10.1145/3577190.3614148

[4] Ighoyota Ben. Ajenaghughrure, Sònia Cláudia Da Costa Sousa, and David Lamas. 2021. 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)*. 000161–000168. https://doi.org/10.1109/SAMI50585.2021.9378655

[5] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. *ACM Trans. Interact. Intell. Syst.* 8, 4, Article 27 (nov 2018), 20 pages. https://doi.org/10.1145/3132743

[6] Severin Bachmann. 2025. Efficient XAI: A Low-Cost Data Reduction Approach to SHAP Interpretability. *J. Artif. Int. Res.* (2025), 21 pages. https://doi.org/10.1613/jair.1.18325

[7] Nattapat Boonprakong, Xiuge Chen, Catherine Davey, Benjamin Tag, and Tilman Dingler. 2023. 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 19 pages. https://doi.org/10.1145/3544548.3580917

[8] John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson. 2016. *Strong Inference in Psychophysiological Science.* Cambridge University Press, 3–15.

[9] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. 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* 1, 2 (2023), 100014. https://doi.org/10.1016/j.chbah.2023.100014

[10] Francesco Chiossi, Ekaterina R. Stepanova, Benjamin Tag, Monica Perusquia-Hernandez, Alexandra Kitson, Arindam Dey, Sven Mayer, and Abdallah El Ali. 2024. PhysioCHI: Towards Best Practices for Integrating Physiological Signals in HCI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 485, 7 pages. https://doi.org/10.1145/3613905.3636286

[11] Peter Christen, David J. Hand, and Nishadi Kirielle. 2023. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Comput. Surv.* 56, 3, Article 73 (Oct. 2023), 24 pages. https://doi.org/10.1145/3606367

[12] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 2 (May 2007), 175–191.

[13] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. arXiv:2101.02235 [cs.CL] https://arxiv.org/abs/2101.02235

[14] Indrajeet Ghosh, Kasthuri Jayarajah, Nicholas Waytowich, and Nirmalya Roy. 2025. Augmenting Personalized Memory via Practical Multimodal Wearable Sensing in Visual Search and Wayfinding Navigation. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*. Association for Computing Machinery, New York, NY, USA, 11–21. https://doi.org/10.1145/3699682.3728340

[15] Ellen R Girden. 1992. *ANOVA: Repeated measures.* Number 84. Sage.

[16] Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science* Volume 6 - 2024 (2025). https://doi.org/10.3389/fcomp.2024.1521066

[17] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. 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* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3531–3540. https://doi.org/10.1145/3485447.3512248

[18] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*. Elsevier, 139–183.

[19] Winston Haynes. 2013. *Benjamini–Hochberg Method.* Springer New York, New York, NY, 78–78. https://doi.org/10.1007/978-1-4419-9863-7_1215

[20] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. Trust in Automation. *IEEE Intelligent Systems* 28, 1 (Jan. 2013), 84–88. https://doi.org/10.1109/MIS.2013.24

[21] Kenneth Holmqvist, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press, United States.

[22] Sarah Jessup, Tamera Schneider, Gene Alarcon, Tyler Ryan, and August Capiola. 2019. *The Measurement of the Propensity to Trust Automation.* 476–489. https://doi.org/10.1007/978-3-030-21565-1_32

[23] Kaixin Ji, Danula Hettiachchi, Flora D. Salim, Falk Scholer, and Damiano Spina. 2024. Characterizing Information Seeking Processes with Multiple Physiological Signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024, Vol. 5)*. ACM, 1006–1017. https://doi.org/10.1145/3626772.3657793

[24] Yucheng Jin, Nyi Nyi Htun, Nava Tintarev, and Katrien Verbert. 2019. ContextPlay: Evaluating User Control for Context-Aware Music Recommendation. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 294–302. https://doi.org/10.1145/3320435.3320445

[25] Sajeev Kunjan, T. S. Grummett, K. J. Pope, D. M. W. Powers, S. P. Fitzgibbon, T. Bastiampillai, M. Battersby, and T. W. Lewis. 2021. The Necessity of Leave One Subject Out (LOSO) Cross Validation for EEG Disease Diagnosis. In *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 558–567. https://doi.org/10.1007/978-3-030-86993-9_50

[26] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[27] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[28] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users. In *Proceedings of Mensch Und Computer 2023* (Rapperswil, Switzerland) *(MuC '23)*. Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/3603555.3603565

[29] Jingshu Li, Yitian Yang, Q. Vera Liao, Junti Zhang, and Yi-Chieh Lee. 2025. As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, New York, NY, USA, Article 1111, 16 pages. https://doi.org/10.1145/3706598.3713336

[30] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. ACM, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182

[31] Q. Vera Liao, Mihaela Vorvoreanu, Hari Subramonyam, and Lauren Wilcox. 2024. UX Matters: The Critical Role of UX in Responsible AI. *Interactions* 31, 4 (jun 2024), 22–27. https://doi.org/10.1145/3665504

[32] Jia Zheng Lim, James Mountstephens, and Jason Teo. 2022. Eye-Tracking Feature Extraction for Biometric Machine Learning. *Front Neurorobot* 15 (Feb. 2022), 796895.

[33] Charles X. Ling, Jin Huang, and Harry Zhang. 2003. AUC: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (Acapulco, Mexico) *(IJCAI'03)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 519–524.

[34] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 183 (April 2024), 39 pages. https://doi.org/10.1145/3641022

[35] Kim Ouwehand, Avalon van der Kroef, Jacqueline Wong, and Fred Paas. 2021. Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales? *Frontiers in Education* Volume 6 - 2021 (2021). https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2021.702616,DOI={10.3389/feduc.2021.702616},ISSN={2504-284X},

[36] Saurin S. Parikh. 2018. Eye Gaze Feature Classification for Predicting Levels of Learning. https://api.semanticscholar.org/CorpusID:53471366

[37] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51, 1 (Feb. 2019), 195–203.

[38] Giuseppe Romeo and Daniela Conti. 2025. Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI Soc.* (July 2025).

[39] Amanda Ross and Victor L. Willson. 2017. *Paired Samples T-Test.* SensePublishers, Rotterdam, 17–19. https://doi.org/10.1007/978-94-6351-086-8_4

[40] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

[41] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422. https://doi.org/10.1145/3581641.3584066

[42] Le Hoang Son. 2016. Dealing with the new user cold-start problem in recommender systems: A comparative review. *Information Systems* 58 (2016), 87–104. https://doi.org/10.1016/j.is.2014.10.001

[43] Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 29 (Oct. 2020), 32 pages. https://doi.org/10.1145/3374218

[44] Xin Sun, Rongjun Ma, Shu Wei, Pablo Cesar, Jos A. Bosch, and Abdallah El Ali. 2026. Understanding trust toward human versus AI-generated health information through behavioral and physiological sensing. *International Journal of Human-Computer Studies* 209 (2026), 103714. https://doi.org/10.1016/j.ijhcs.2025.103714

[45] Xin Sun, Shu Wei, Jos A Bosch, Isao Echizen, Saku Sugawara, and Abdallah El Ali. 2026. Seeing the Reasoning: How LLM Rationales Influence User Trust and Decision-Making in Factual Verification Tasks. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26)*. Association for Computing Machinery, New York, NY, USA, Article 585, 7 pages. https://doi.org/10.1145/3772363.3798613

[46] Anna Taudien, Andreas Fügener, Alok Gupta, and Wolfgang Ketter. 2022. The Effect of AI Advice on Human Confidence in Decision-Making. https://doi.org/10.24251/HICSS.2022.029

[47] Tobii AB. 2024. Tobii Pro Lab (Version 1.2xx).

[48] Dereck Toker and Cristina Conati. 2017. Leveraging Pupil Dilation Measures for Understanding Users' Cognitive Load During Visualization Processing. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 267–270. https://doi.org/10.1145/3099023.3099059

[49] Ralf Van der Lans, Michel Wedel, and Rik Pieters. 2011. Defining eye-fixation sequences across individuals and tasks: the Binocular-Individual Threshold (BIT) algorithm. *Behavior research methods* 43 (2011), 239–257.

[50] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. arXiv:2212.06823 [cs.HC] https://arxiv.org/abs/2212.06823

[51] Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. 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* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 28, 14 pages. https://doi.org/10.1145/3613904.3642018

[52] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (Oct. 2021), 39 pages. https://doi.org/10.1145/3476068

[53] L Wang and J A Stern. 2001. Saccade initiation and accuracy in gaze shifts are affected by visual stimulus significance. *Psychophysiology* 38, 1 (Jan. 2001), 64–75.

[54] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 109–116. https://doi.org/10.1109/HRI.2016.7451741

[55] Xujinfeng Wang, Yicheng Yang, Da Tao, and Tingru Zhang. 2023. The impact of AI transparency and reliability on human-AI collaborative decision-making. In *AHFE International*. AHFE International.

[56] Michel Wedel, Rik Pieters, and Ralf van der Lans. 2023. Modeling eye movements during decision making: A review. *Psychometrika* 88, 2 (June 2023), 697–729.

[57] Wikipwdia. 2023. Hallucination (artificial intelligence) - Wikipedia. https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence). (Accessed on 10/10/2023).

[58] Jiayi Yao, Haibo Sun, and Nianwen Xue. 2025. Fact-checking AI-generated news reports: Can LLMs catch their own lies? arXiv:2503.18293 [cs.CL] https://arxiv.org/abs/2503.18293

[59] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making *(FAT* '20)*. ACM, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[60] Ömer Sümer, Efe Bozkir, Thomas Kübler, Sven Grüner, Sonja Utz, and Enkelejda Kasneci. 2021. FakeNewsPerception: An eye movement dataset on the perceived believability of news stories. *Data in Brief* 35 (2021), 106909. https://doi.org/10.1016/j.dib.2021.106909

[61] Jakub Štěpán Novák, Jan Masner, Petr Benda, Pavel Šimek, and Vojtěch Merunka. 2024. Eye Tracking, Usability, and User Experience: A Systematic Review. *International Journal of Human–Computer Interaction* 40, 17 (2024), 4484–4500. https://doi.org/10.1080/10447318.2023.2221600 arXiv:https://doi.org/10.1080/10447318.2023.2221600