ThoughtTrace: UNDERSTANDING USER THOUGHTS IN REAL-WORLD LLM INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conversational AI has now reached billions of users, yet existing datasets capture only what people *say*, not what they *think*. We introduce *ThoughtTrace*, the first large-scale dataset that pairs real-world multi-turn human–AI conversations with users' self-reported thoughts: their *reasons* for sending prompts and *reactions* to assistant responses. *ThoughtTrace* comprises 1,058 users, 2,155 conversations, 17,058 turns, and 10,174 thought annotations collected across 20 language models. Our analysis shows that *ThoughtTrace* captures long-horizon, topically diverse interactions, and that thoughts are semantically distinct from messages, difficult for frontier LLMs to infer from context, diverse in content, and tied to conversation stages. We further demonstrate the utility of thoughts for downstream modeling. First, thoughts improve user-behavior prediction as inference-time context. Second, thought-guided rewrites provide fine-grained alignment signals for training personalized assistants. Together, *ThoughtTrace* establishes user thoughts as a new data modality for studying the cognitive dynamics behind human–AI interaction and provides a foundation for building assistants that better understand and adapt to users' latent goals, preferences, and needs.

1 INTRODUCTION

Conversational AI systems have now been deployed at an unprecedented scale, processing billions of user interactions every day. While extensive work focuses on what users *say* during these interactions (Zheng et al., 2023; Zhao et al., 2024b; Baumann et al., 2026; Jin et al., 2025; Shi et al., 2024), understanding what users actually *think* during the conversations remains a largely unexplored dimension of human-AI interaction.

User thoughts are the unspoken cognitive context behind each message: the motivation and goal driving the request, the context and constraints grounding it, the content or style expectations for the response, and the interpretations and reactions to the assistant's reply. Figure 1 illustrates why this hidden layer matters. The observed initial user message about preparing for a trip reads as a generic travel query, but the unobservable *thought* exposes the anxiety of an inexperienced international traveler. After the assistant replies with a standard checklist, the user's thought reveals dissatisfaction that the next message never explicitly states: the response feels generic and overlooks the conference context. The user's follow-up message operationalizes this private reaction by requesting a structured breakdown. Capturing these thoughts and their dynamics closes the gap between observable utterances and hidden user intents, providing richer signals for training and evaluation.

We introduce *ThoughtTrace*, the first framework and dataset for understanding user thoughts during real-world human-AI interactions at scale. By asking users to engage in natural conversations while articulating contextually grounded thoughts, we collect a rich corpus of first-person cognitive traces that illuminate the lived experience of interacting with AI systems.

ThoughtTrace features high-quality, long-horizon interactions grounded in open-ended real-world tasks performed by a diverse user base: 1,058 users, 2,155 timestamped conversations, 17,058 interaction turns, and 10,174 thought annotations, collected via a chatbot service powered by 20 different language models. Each conversation includes: (1) naturalistic multi-turn dialogue between a user and an AI assistant; (2) user-reported thoughts aligned to individual user and assistant messages, including *reasons* for sending messages and *reactions* to assistant responses; (3) post-task descriptions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

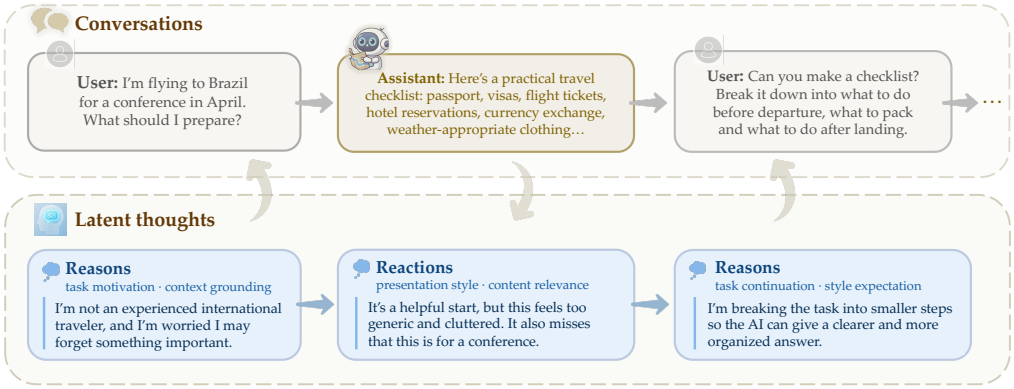


Figure 1: **A representative example from *ThoughtTrace*.** A user interacts with a chatbot to complete daily tasks through multi-turn conversations (top), while annotating their latent thoughts during the conversations (bottom). Thoughts take two forms: *reasons* for sending user prompts and *reactions* to assistant responses, which can be categorized into several types (e.g., *task motivation*, *style expectation*). Latent thoughts reveal users' thought traces that drive the human-AI interactions in multi-turn conversations, providing valuable signals for user modeling and improving AI assistance.

of what users completed and what they expected from the AI; and (4) user demographic information such as age, gender, education level, and occupation.

Our analysis highlights the properties and utility of thoughts along three axes: (1) **Conversation properties** (Section 4.1): *ThoughtTrace* features representative users, long-horizon conversations, broad topical coverage, and frequent extensions across turns. (2) **Thought properties** (Section 4.2): thoughts differ from messages, are difficult for frontier LLMs to infer, are diverse in content, and are tied to conversation stages. (3) **Thought utility** (Section 5): thoughts predict user behavior during inference (+32.4% relative gain), and provide fine-grained alignment signals (+25.6% win rate).

ThoughtTrace opens several directions for future research. On **user modeling**, it enables systematic study of what users think during conversations, how context shapes these thoughts, how thoughts shape user utterances, and how these dynamics vary across demographic groups. On **model training**, user thoughts provide a new supervisory signal that models can predict, learn from, and align with, offering a path toward assistants that better capture users' latent goals, expectations, and reactions. On **evaluation**, *ThoughtTrace* enables benchmarks for thought prediction and supports thought-centered measures of user satisfaction, moving evaluation beyond surface-level utterances toward latent intent and subjective experience.

Our contributions are summarized as follows: (1) We introduce *thoughts* as a new data modality for human-AI interaction research, and release *ThoughtTrace*, a large-scale dataset pairing naturalistic multi-turn conversations with rich thought annotations and demographic metadata. (2) We characterize the conversational and cognitive structure of *ThoughtTrace* along multiple axes, showing that thoughts are latent, hard to infer, diverse, and stage-dependent. (3) We demonstrate the utility of thoughts for predicting user behavior and aligning language models.

2 Related Work

Real-World Human-AI Conversations. There have been recent datasets of real-world human-AI conversations, including general chat datasets such as WildChat (Zhao et al., 2024b) and LMSYS-Chat-1M (Zheng et al., 2023) and domain-specific datasets such as SWE-Chat (Baumann et al., 2026) for software engineering. Additionally, PRISM (Kirk et al., 2024) paired conversation logs with sociodemographic surveys and stated preferences. Building on such corpora, recent works have developed methods to effectively extract supervisory signals such as satisfaction cues from natural conversations (Zhao et al., 2024a; Shi et al., 2024; Jin et al., 2025; Peng et al., 2026; Buening et al., 2026). Across these efforts, any view of the user is limited to what they explicitly verbalize; even PRISM elicits only ratings or stated preferences over outputs, leaving the intents and thought processes behind their messages unobserved. *ThoughtTrace* addresses this gap by pairing real conversations with underlying thought dynamics self-reported by the users.

User Thoughts. There has been increasing interest in machine Theory of Mind (ToM) (Wimmer & Perner, 1983), the ability to infer people's latent mental states from their behavior. However, most work focuses on structured ToM reasoning (Baker et al., 2009; 2017), where inferences are limited to a few well-defined mental variables, such as goals, beliefs, and desires, grounded in simple contexts (Ullman, 2023; Kim et al., 2023; Shapira et al., 2024; Jin et al., 2024; Shi et al., 2025; Fan et al., 2025; Sclar et al., 2023; Zhang et al., 2025; Jha et al., 2024), and thus fails to capture the dynamics of latent thoughts during interactions (see Appendix C.1 for how thoughts differ from these variables). While recent work leverages dynamic mental state inference to enhance AI assistance (Zhang et al., 2025; Zhou et al., 2025; Zhang et al., 2026), systematic analysis and large-scale collection of user thoughts in human-AI interactions remain lacking. *ThoughtTrace* provides a new paradigm for collecting and analyzing users' latent thoughts during multi-turn human-AI conversations.

User Simulations. There has been increasing interest in building user simulators for training and evaluating AI assistants to address the data gap (Qian et al., 2025; Park et al., 2024; Wu et al., 2026; Binz et al., 2025; Naous et al., 2025; Park et al., 2023; Abdulhai et al., 2025). To do so, these works have heavily relied on prompting LLMs (Park et al., 2024; Piao et al., 2025) or finetuning LLMs on ground-truth responses or persona-consistent behavior (Binz et al., 2025; Naous et al., 2025; Kolluri et al., 2025; Abdulhai et al., 2025; Mehri et al., 2025; Zhu et al., 2025). However, recent works have found that existing simulators are biased and unfaithful (Zhou et al., 2026; Seshadri et al., 2026). While HumanLM (Wu et al., 2026) attempts to mitigate this by aligning simulated user conversations with users' internal states, its training still relies on synthetic user thoughts due to the lack of real thought data. The first-person thought traces from real users in real interactions in *ThoughtTrace* may provide valuable data for training more realistic user simulators.

3 Data Collection

3.1 What are *Thoughts*?

Thoughts refer to users' latent cognitive context in human–AI conversations. Unlike observable utterances, which are often lossy representations of intent due to the principle of least effort (Zipf, 2016), thoughts capture the unspoken mental content that motivates those utterances. Because thoughts are richer and faster-moving than verbalized language, conversations can transmit only a fraction of them in real time. Conversational language is also shaped by pragmatic and utility-driven pressures: speakers produce utterances that are efficient, socially appropriate, and goal-directed, rather than fully transparent reflections of their mental states (Sperber & Wilson, 1986).

As shown in Figure 1, in our data collection, thoughts are annotated as either *reactions*, which reflect how users internally respond to an assistant message, or *reasons*, which explain why users send a particular message. We collect both types at each turn because they jointly shape how users proceed in the next turn. Specifically, reactions indicate how users perceive the model, while reasons reveal how users want the model to understand their needs and preferences. Thoughts also provide a taxonomy of users' mental processes complementary to standard Theory of Mind (ToM) variables such as goals, beliefs, and desires. In particular, thoughts are tied to specific turns of the interaction, characterizing a user's dynamic reasoning process. A single thought may span multiple ToM variables while revealing information none of them captures (see Appendix C.1 for more discussion).

3.2 Methodology

We recruited participants via Prolific and redirected them to our data collection platform to complete trials following the procedure below. This study was approved by an institutional review board.

Step 1: User consent. Participants provided informed consent acknowledging voluntary participation, guaranteed anonymity, and the right to withdraw at any time.

Step 2: Tutorial and quiz. Participants first completed a guided tutorial introducing the chat interface and demonstrating how to send messages, annotate thoughts, start a new chat, and finish a task. They must then pass a short comprehension quiz before proceeding.

Step 3: Conversations with thoughts. Participants completed two open-ended, self-defined tasks, each within a 10-minute window, while chatting naturally with the AI and privately annotating their

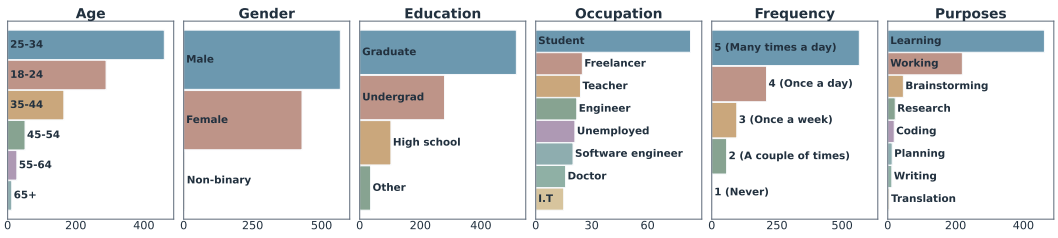


Figure 2: **Participant demographics and AI usage patterns in *ThoughtTrace*.** The dataset covers age, gender, education level, occupation, frequency of AI usage, and primary purposes for using AI.

reasons for sending each message and their reactions to each assistant response. Each task could span multiple multi-turn conversations: participants were free to start a new conversation or end the task at any time, mirroring real-world use of conversational AI systems. Participants could add a reason and/or reaction whenever they wished. This (1) avoids forcing them to fabricate a thought annotation when they don't have a clear one and (2) minimizes interruption to the natural interaction. Overall, 58.8% of turns received at least one thought annotation (see Appendix C.5 for details).

Step 4: Survey. After each task, participants described what they completed and what they expected from the AI. After both tasks, they filled out a demographic survey covering age, gender, education, occupation, AI usage frequency, and primary purposes.

Details of the data collection methods, platform design, and limitations are provided in Appendix C.

3.3 Models Used

Each participant interacted with one of 20 different models. We included frontier models available at the time of the study (e.g., GPT-5.4, Gemini 3.1 Pro Preview, Grok 4.20, and Opus 4.6), as well as smaller, open-weight models for comparison. Users were unaware of which model they were interacting with. Detailed statistics for each model are provided in Appendix A.

4 Data Properties

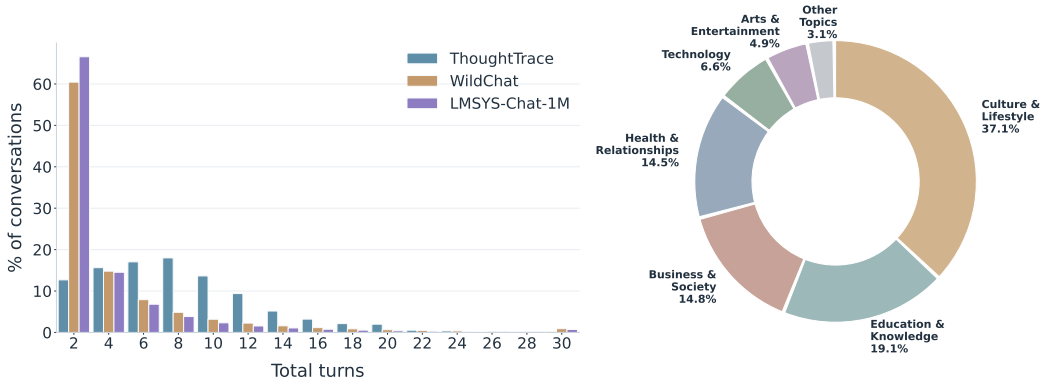
We characterize the data in *ThoughtTrace* along two complementary axes: (1) properties of the conversations (Section 4.1) and (2) properties of the thoughts that drive the conversations (Section 4.2).

4.1 Properties of Conversations

We highlight three conversation-level properties: a representative user base, long-horizon and topically diverse interactions, and the dominance of conversational turns that extend prior tasks.

[Conversation Property 1] *ThoughtTrace* captures a representative spectrum of users. In Figure 2, we summarize the responses to our background survey (details in Appendix C.6). Unlike existing in-the-wild conversation datasets such as WildChat (Zhao et al., 2024b), which contain little participant-level information, *ThoughtTrace* pairs each conversation with rich demographic and usage metadata. Overall, the sample spans a broad range of backgrounds: participants range from 18 to 65+ in age, cover multiple education levels, and represent a variety of occupations, including students, freelancers, teachers, engineers, and others. That said, the participant distribution is skewed towards the 18–34 age range and those with at least an undergraduate degree, broadly consistent with the demographic profile of frequent generative AI users (Liu & Wang, 2026; Bick et al., 2026). Most participants report frequent AI use, often one or more times per day, for a range of purposes. The most common uses are learning and working, followed by brainstorming, research, and coding.

[Conversation Property 2] *ThoughtTrace* features long-horizon diverse conversations. We compute conversation lengths at both the turn and token levels, with implementation details in Appendix D.2. As shown in Figure 3(a), *ThoughtTrace* exhibits a substantially more balanced turn distribution, peaking around 6–8 turns with a median of 8 turns, whereas WildChat and LMSYS-Chat-1M are heavily skewed toward short 2-turn exchanges, which alone account for about 60% and 67% of their conversations, respectively. The cumulative token distribution per conversation follows a similar trend (Appendix B.6). This long-horizon property is critical because real-world AI usage



(a) Turn distribution across the three datasets.

(b) Topic distribution in *ThoughtTrace*.

Figure 3: ***ThoughtTrace* covers long-horizon, topically diverse conversations.** (a) Turn distribution comparison among *ThoughtTrace*, WildChat, and LMSYS-Chat-1M: *ThoughtTrace* peaks at 6–8 turns, while the baselines skew heavily toward 2-turn exchanges. (b) Distribution of conversation topics in *ThoughtTrace*, grouped into seven broad domains, with no single category dominating.

is increasingly shifting toward sustained multi-turn interactions such as iterative coding, research, and planning, where tasks are more complex and users' underlying intentions evolve across turns rather than being captured in a single prompt.

To characterize topical coverage, we label the relevant topics of each conversation, with implementation details in Appendix D.2. Conversations are distributed across seven broad categories (Figure 3(b)) and 36 fine-grained subtopics (see Figure 13 in Appendix B.7 for the full breakdown). Culture & Lifestyle is the most prevalent broad topic category (covering areas such as travel, dining, and daily life), while Education & Knowledge as well as Business & Society are also well represented. At the fine-grained level, nine subtopics each exceed 5% of all topic labels (spanning Travel, Lifestyle, Food, Business, Geography, Education, Relationships, Health, and Technology), with a long tail of more specialized topics covering the remaining share. We also collect participants' task descriptions and AI expectations, with details in Appendix C.6 and visualizations in Appendix B.8.

[Conversation Property 3] *ThoughtTrace* conversations are dominated by task extension. We analyze conversational structure by labeling the multi-turn relationship of each user message into one of five types: (1) First request (**25.2%**); (2) Completely new request (**12.5%**); (3) Re-attempt/revision on prior task (**2.9%**); (4) New variation of prior task (**2.3%**); and (5) Extend, deepen, or build on prior task (**57.0%**). Implementation details are provided in Appendix D.3, and the overall distribution is shown in Figure 16. Figure 4 visualizes how these relationships transition across the first three user turns. Extension dominates from turn 2 onward and becomes increasingly prevalent in later turns. Re-attempts and variations occur infrequently throughout, suggesting that users rarely need to rephrase or retry their requests.

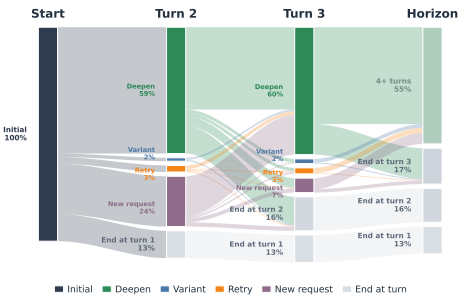


Figure 4: **Multi-turn relationship flow.** Turn-to-turn transitions of relationship labels across the first three turns and beyond, showing how conversations evolve from the initial request.

4.2 Properties of Thoughts

We highlight four thought-level properties: thoughts are different from messages, difficult for frontier LLMs to infer, span diverse reason and reaction categories, and are tied to conversation stages.

[Thought Property 1] Thoughts are different from messages. Do thoughts merely restate users' messages, or capture new information? We first evaluate at the **embedding** level: Figure 5 visualizes the pairwise embedding differences between (i) a *user message* and the *next user message*, (ii) a *user message* and its corresponding *reason*, and (iii) a user's *reaction* to an assistant response and their

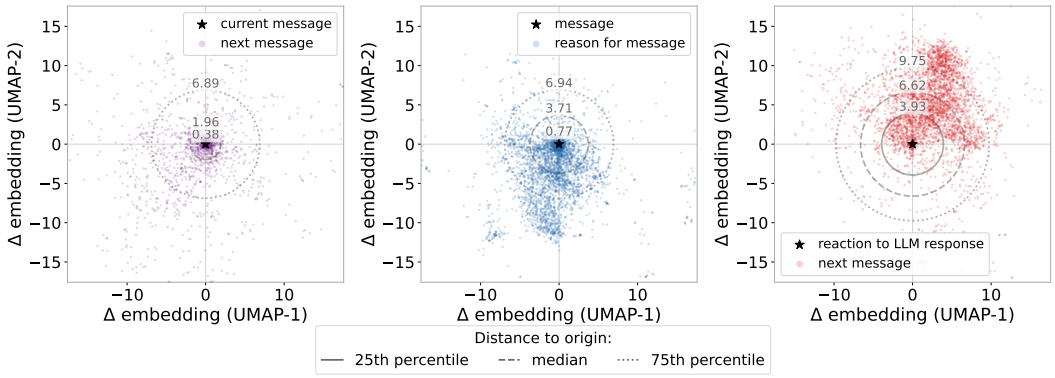


Figure 5: **UMAP projections of embedding differences across three paired settings.** The star denotes the reference text embedding, and each dot represents the paired text embedding. Distance from the origin reflects the magnitude of the semantic shift between the paired texts. Circle annotations denote the 25th, 50th, and 75th percentile distances from the origin.

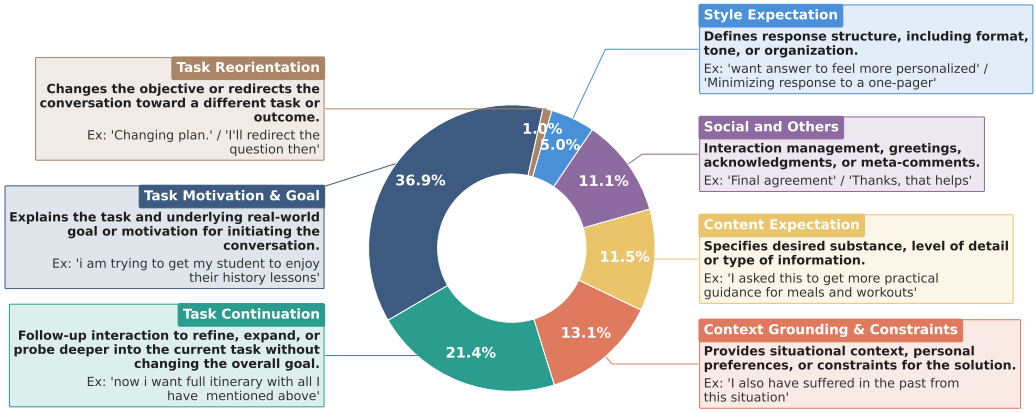


Figure 6: **Distribution of seven user reason types in *ThoughtTrace*, with definitions and examples from the dataset.** Task Motivation & Goal is the most prevalent (36.9%), followed by Task Continuation (21.4%) and Context Grounding & Constraints (13.1%).

following *next user message*. Consecutive user messages remain semantically close, reflecting the local coherence of conversation, whereas message–reason pairs show larger distances and reaction–next-message pairs exhibit the widest dispersion; quantitative distributional metrics in Appendix B.9 confirm the same trend. We then measure **semantic coverage via an LLM-based judge**, scoring on a 1 (no overlap) to 5 (full coverage) rubric how well a *user message* covers (i) its *reason* and (ii) the *reaction* to the prior assistant response (see Appendix D.4 for implementation details). Average scores are 3.22 for reasons (partial overlap, missing the core of the thought) and 2.00 for reactions (minimal overlap). A **human study** with 30 independent raters recruited on Prolific further corroborates this: on a 1–5 Likert scale, they agreed that thoughts contain information beyond the messages (4.0) and help them better understand user behavior (4.2) (Appendix B.1). Together, these results show that thoughts capture substantial latent information not verbalized in conversation.

[Thought Property 2] Thoughts are difficult for LLMs to infer. We prompt LLMs to infer (1) the user's reason for their most recent message, given the conversation up to that point, and (2) the user's reaction to the assistant's most recent message, given the conversation up to that point plus the user's next message if available. An LLM-as-a-judge scores each inference against the human annotation on a 1–5 semantic similarity scale. Implementation details are provided in Appendix D.5. Averaged across three frontier models (GPT-5.4, Gemini 3.1 Pro Preview, and Claude Opus 4.6), the mean similarity score is 2.93 for reasons (2.83, 3.02, 2.94, respectively) and 2.54 for reactions (2.36, 2.87, 2.40), all falling near or below partial overlap (3). The gap reflects the fact that thoughts are underspecified by surface-form text: multiple plausible reasons or reactions are consistent with the same context, and the correct one often depends on unobservable constraints, stakes, or interpreta-

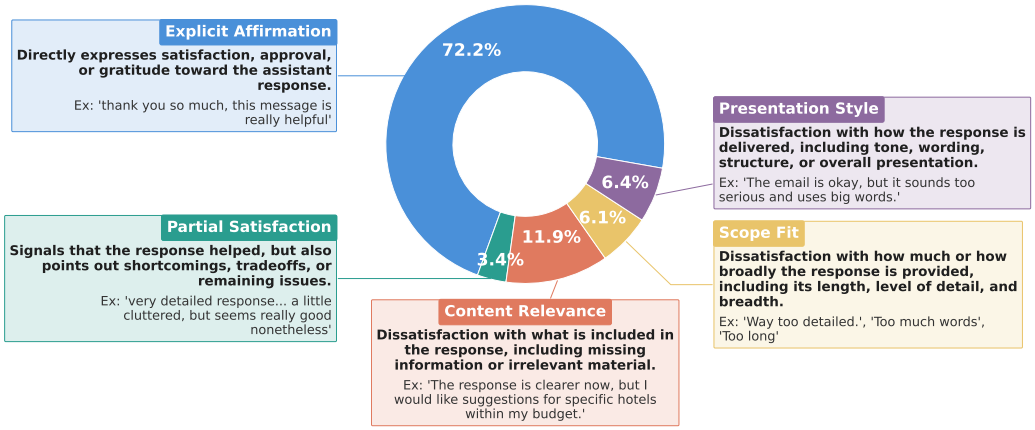
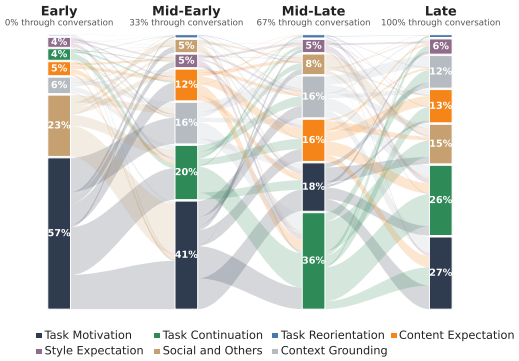
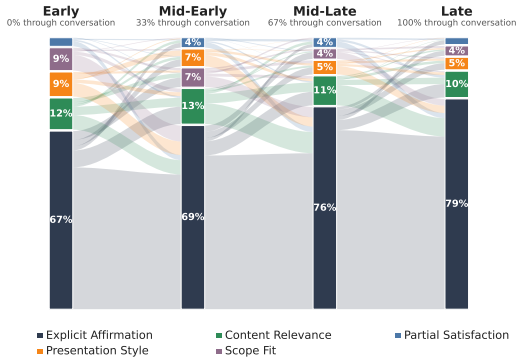


Figure 7: **Distribution of five user reaction types in *ThoughtTrace*, with definitions and examples from the dataset.** Explicit Affirmation dominates (72.2%), while dissatisfaction is often driven by Content Relevance (11.9%), Presentation Style (6.4%), and Scope Fit (6.1%).



(a) Reason types across conversation stages.



(b) Reaction types across conversation stages.

Figure 8: **Thought dynamics across conversation stages.** (a) Reason-type distribution shifts from Task Motivation & Goal in early turns to Task Continuation and context- and expectation-driven reasons in later stages. (b) Reaction-type distribution shows a steady increase in Explicit Affirmation from early to late stages.

tions from users. Appendix B.2 shows qualitative failure cases in which models misread the user's underlying intent or fabricate reactions the users did not have.

[Thought Property 3] Thoughts are diverse in content. To analyze this diversity, we label user thoughts using an LLM-based annotation framework (details in Appendix D.6). As shown in Figure 6, the reasons behind user utterances span seven distinct categories, ranging from high-level drivers such as *Task Motivation & Goal* (36.9%) and *Task Continuation* (21.4%) to finer-grained context and preference specifications such as *Context Grounding & Constraints* (13.1%), *Content Expectation* (11.5%), and *Style Expectation* (5.0%). Complementing this, Figure 7 shows that user reactions decompose into five categories: while *Explicit Affirmation* dominates at 72.2% and *Partial Satisfaction* accounts for 3.4%, the remaining reactions reveal targeted sources of dissatisfaction, including *Content Relevance* (11.9%), *Presentation Style* (6.4%), and *Scope Fit* (6.1%).

[Thought Property 4] Thought dynamics depend on conversation stages. Figure 8a shows that *Task Motivation & Goal* dominates early turns, while *Task Continuation* increases and becomes the primary driver in the Mid-Late stage. *Context Grounding & Constraints* and expectation-related reasons remain a substantial portion throughout the middle stages. Figure 8b shows a parallel shift in reactions: *Explicit Affirmation* increases from 67% in early stages to 79% in later stages, while more critical reactions such as *Presentation Style* and *Scope Fit* decline, suggesting that user satisfaction improves as interactions converge toward acceptable responses. Figure 17 corroborates these trends at the message-relationship level and further shows that users predominantly extend the

conversation regardless of their annotated reaction type. By contrast, thought types exhibit no clear relationship with conversation topics or lengths, with additional results discussed in Appendix B.10 and Figures 18–21. Implementation details are in Appendix D.7.

5 Utility of Thoughts

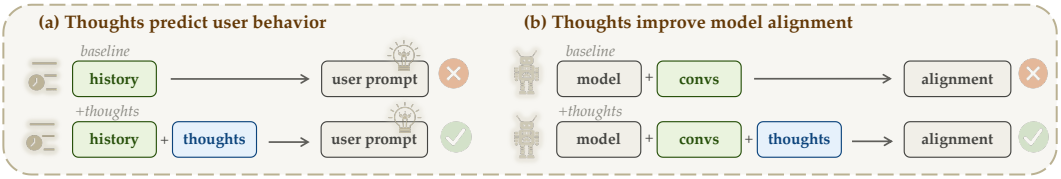


Figure 9: **Two experiments demonstrating the utility of thoughts.** Thoughts provide actionable signals for (a) predicting user behavior and (b) improving model alignment.

We define the utility of thoughts as the actionable signals they provide beyond what is observable in conversation transcripts. As shown in Figure 9, we validate this utility through two experiments: predicting user behavior (Section 5.1) and improving model alignment (Section 5.2), pointing toward future work on user modeling, thought-centered evaluation, and personalized assistant training.

5.1 Thoughts Predict User Behavior

Predicting user behavior is important as (1) it helps models anticipate user needs and provide more proactive, personalized assistance; and (2) it supports high-fidelity user simulators, which provide a scalable and reproducible alternative to real human interaction during model training and evaluation.

Experimental setup. We test whether access to thought annotations at inference time improves the LLM's ability to anticipate the user's next message. For each conversational turn from *ThoughtTrace*, we compare two settings: (1) predicting the next message from the conversation history alone, and (2) predicting it from the same history augmented with the user's annotated reasons and reactions. We evaluate six frontier models (GPT-5.4, Gemini 3.1 Pro Preview, Claude Opus 4.6, Kimi-K3, GLM-5.2, Muse Spark 1.1) under both settings, and score each prediction's semantic similarity to the ground truth on a 0–100 scale using an LLM judge randomly drawn from the other models. Details are in Appendix D.8.

Table 1: **User message prediction results.** Six frontier models are evaluated, with and without access to annotated thoughts at inference time. Subscripts denote absolute gains over History-only.

Method	GPT-5.4	Gemini 3.1 Pro Preview	Claude Opus 4.6	Kimi-K3	GLM-5.2	Muse Spark 1.1	Avg.
History-only	21.4	22.1	21.3	25.9	33.3	31.5	25.9
Thought-augmented	**27.4**$_{+6.0}$	**28.9**$_{+6.8}$	**35.5**$_{+14.2}$	**36.2**$_{+10.3}$	**37.8**$_{+4.5}$	**39.8**$_{+8.3}$	**34.3**$_{+8.4}$

Results. As shown in Table 1, access to thought annotations substantially improves next-message prediction across all six models, raising the average performance from 25.9 to 34.3, a 32.4% relative gain. The effect is largest for Claude Opus 4.6 and Kimi-K3, whose performance increases by 14.2 and 10.3, respectively, while the remaining models show smaller but consistent gains. These results suggest that thoughts provide actionable signals beyond the observable conversation history. Qualitative examples are in Appendix B.3.

Implications for future research. Our results demonstrate that user thoughts can steer user behavior predictions, which suggests the value of thoughts in simulating users. For example, unlike prior work that trains user simulators by predicting the *next user message* (Naous et al., 2025; Abdulhai et al., 2025; Wu et al., 2026), future work could train models to jointly predict *thoughts* and user messages. Strong user simulators can help anticipate user needs and thereby guide models to assist users in a more proactive and personalized manner (Qian et al., 2025; Sun et al., 2025).

5.2 THOUGHTS IMPROVE MODEL ALIGNMENT

Model alignment helps models better match human intentions, values, and preferences, and real user interactions provide natural, multifaceted signals for it.

Experimental setup. Prior work on learning from natural conversations revises unsatisfactory responses using users' follow-up messages, pairing these *message-guided rewrites* with the original responses for preference learning (Shi et al., 2024; Jin et al., 2025). Leveraging thoughts in *ThoughtTrace*, we instead identify unsatisfactory responses via the *dissatisfaction reaction labels* from Section 4.2 and prompt the model to revise them using the *thought content*, producing *thought-guided rewrites*. Both are paired with originals for DPO training (Rafailov et al., 2023). We compare: (1) the base Qwen3.5-4B (Yang et al., 2025); (2) message-guided rewrites on WildChat; (3) message-guided rewrites on *ThoughtTrace*; and (4) thought-guided rewrites on *ThoughtTrace*. Models are evaluated on Arena-Hard (Li et al., 2024), a robust instruction-following benchmark with 98.6% correlation to human preference. Details are in Appendix D.9.

Results. As shown in Table 2, fine-tuning Qwen3.5-4B on *ThoughtTrace* substantially improves Arena-Hard performance, with thought-guided rewrites achieving the largest style-controlled gains over both the base model (+25.6%) and the WildChat baseline (+6.6%). We highlight three findings: **(1)** within *ThoughtTrace*, thought-guided rewrites outperform message-guided ones (+4.5%), indicating that thoughts encode richer dissatisfaction and revision signals than users explicitly articulate in messages, and human annotators prefer thought-guided rewrites in 70.0% of comparisons (Appendix B.4); **(2)** across the same *ThoughtTrace* conversations, thoughts surface 1,000 dissatisfaction instances compared to 450 in messages (2.2× more), yielding denser supervision, while thoughts still outperform messages at a matched size of 450 (Appendix B.5); and **(3)** compared to the WildChat baseline, the message-guided variant of *ThoughtTrace* uses fewer conversations and a smaller training set yet still outperforms it (+2.1%), reflecting the higher quality of *ThoughtTrace*. More broadly, thoughts provide ground-truth user reactions rather than behavioral proxies, and unify *which* response is unsatisfactory and *how* to revise it into a single supervision signal.

Table 2: **Model alignment results on Arena-Hard.** We report both win rates (%) and style-controlled win rates (SC Win, %) with 95% confidence intervals.

Method	Win	SC Win
Qwen3.5-4B	24.6±2.6	22.5±2.7
+ WildChat	41.8±3.2	41.5±3.5
+ *ThoughtTrace* (messages)	44.0±3.0	43.6±3.0
+ *ThoughtTrace* (thoughts)	**47.9**±3.2	**48.1**±3.6

Implications for future research. We advocate broader adoption of our framework for collecting thoughts as richer and more effective signals for model training. In terms of training methods, our experiments use only reactions; future work may additionally incorporate reasons and leverage both signals jointly. Moreover, thought-guided supervision could be extended to reward modeling and online alignment (Peng et al., 2026), and thought-guided On-Policy Distillation (OPD) may provide rich signals for online improvement (Wang et al., 2026; Buening et al., 2026; Hübotter et al., 2026).

6 CONCLUSION

We introduce *ThoughtTrace*, the first large-scale dataset that pairs real-world human-AI conversations with users' self-reported thoughts. Our analysis establishes thoughts as a distinct data modality: they capture latent information beyond surface messages, are difficult for frontier LLMs to infer, span diverse content, and vary across conversation stages. We further demonstrate their utility, showing that thoughts improve user behavior prediction at inference time and provide fine-grained alignment signals for training. Together, these results position user thoughts as a foundational signal for studying human-AI interaction and building assistants that better model and learn from users.

Limitations and Future Work. *ThoughtTrace* has several limitations inherent to in-situ thought collection (Appendix C.12). First, asking users to externalize thoughts may shape the interaction itself, as anticipating annotation can sharpen or polarize their reasoning. Second, the dataset captures only consciously accessible reasoning, leaving subconscious judgments unobserved. Third, recruitment through Prolific introduces a modest selection effect, though our demographic analysis suggests the sample remains broadly representative of frequent AI users. Finally, we cover only two downstream use cases, and a more comprehensive empirical investigation is left to future work.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. We did not use generative AI tools for retrieval and discovery, research ideation or execution, drafting sections of the paper, generating synthetic datasets, or proving mathematical claims. We manually reviewed and edited all AI-assisted writing to ensure accuracy, clarity, and consistency with the authors' intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All studies in this work that involve human subjects were approved by an institutional review board. Participants provided informed consent acknowledging voluntary participation, guaranteed anonymity, and the right to withdraw at any time. We take care to protect participant privacy in the collection, analysis, and release of the data.

REPRODUCIBILITY STATEMENT

We include the *ThoughtTrace* dataset in the Supplementary Material. The paper and appendix provide details on the data collection procedure, analysis methods, experimental setup, prompts, training, and evaluation. We will release all code used for data analysis and experiments to support reproduction of our results.

REFERENCES

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*, 2025.

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 0064, 2017.

Joachim Baumann, Vishakh Padmakumar, Xiang Li, John Yang, Diyi Yang, and Sanmi Koyejo. Swe-chat: Coding agent interactions from real users in the wild. *arXiv preprint arXiv:2604.20779*, 2026.

Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative ai. *Management Science*, 2026.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Thomas Kleine Buening, Jonas Hübotter, Barna Pásztor, Idan Shenfeld, Giorgia Ramponi, and Andreas Krause. Aligning language models from user interactions. *arXiv preprint arXiv:2603.12273*, 2026.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. Somitom: Evaluating multi-perspective theory of mind in embodied social interactions. *arXiv preprint arXiv:2506.23046*, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Kunal Jha, Tuan Anh Le, Chuanyang Jin, Yen-Ling Kuo, Joshua B Tenenbaum, and Tianmin Shu. Neural amortized inference for nested multi-agent reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 530–537, 2024.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, 2024.

Chuanyang Jin, Jing Xu, Bo Liu, Leitian Tao, Olga Golovneva, Tianmin Shu, Wenting Zhao, Xian Li, and Jason Weston. The era of real-world human interaction: Rl from user conversations. *arXiv preprint arXiv:2509.25137*, 2025.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, 2023.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30084–30099, 2025.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Yan Liu and He Wang. Who on earth is using generative ai? *World Development*, 199:107260, 2026.

Shuhaib Mehri, Xiaocheng Yang, Takyoung Kim, Gokhan Tur, Shikib Mehri, and Dilek Hakkani-Tür. Goal alignment in llm-based user simulators for conversational ai. *arXiv preprint arXiv:2507.20152*, 2025.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. *arXiv preprint arXiv:2510.06552*, 2025.

OpenAI. text-embedding-3-large. `https://developers.openai.com/api/docs/models/text-embedding-3-large`, 2024. Accessed: 2026-04-28.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52, 2024.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Lei Hou, and Juanzi Li. Wildreward: Learning reward models from in-the-wild human interactions. *arXiv preprint arXiv:2602.08829*, 2026.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. 2025.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Jielin Qiu, Zhiwei Liu, Haolin Chen, Shirley Kokane, Heng Ji, Weiran Yao, Shelby Heinecke, et al. Userrl: Training interactive user-centric agent via reinforcement learning. *arXiv preprint arXiv:2509.19736*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13960–13980, 2023.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. Lost in simulation: Llm-simulated users are unreliable proxies for human users in agentic evaluations. *arXiv preprint arXiv:2601.17087*, 2026.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2257–2273, 2024.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1510–1519, 2025.

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Jauhar, Sihao Chen, Shan Xia, et al. Wildfeedback: Aligning llms with in-situ user interactions and feedback. *arXiv preprint arXiv:2408.15549*, 2024.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. Training proactive and personalized llm agents. *arXiv preprint arXiv:2511.02208*, 2025.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Yinjie Wang, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. Openclaw-rl: Train any agent simply by talking. *arXiv preprint arXiv:2603.10165*, 2026.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. Humanlm: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunchi Zhang, Jin Lu, Chuanyang Jin, Yichao Zhou, Zhining Zhang, and Tianmin Shu. Mindzero: Learning online mental reasoning with zero annotations. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. Autotom: Scaling model-based mental inference via automated agent modeling. *arXiv preprint arXiv:2502.15676*, 2025.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, et al. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *arXiv preprint arXiv:2407.17468*, 2024a.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023.

Xuhui Zhou, Valerie Chen, Zora Zhiruo Wang, Graham Neubig, Maarten Sap, and Xingyao Wang. Tom-swe: User mental modeling for software engineering agents. *arXiv preprint arXiv:2510.21903*, 2025.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, et al. Mind the sim2real gap in user simulation for agentic tasks. *arXiv preprint arXiv:2603.11245*, 2026.

Jian-Qiao Zhu, Hanbo Xie, Dilip Arumugam, Robert C Wilson, and Thomas L Griffiths. Using reinforcement learning to train large language models to explain human decisions. *arXiv preprint arXiv:2505.11614*, 2025.

George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.

A Details of Models Used in *ThoughtTrace*

ThoughtTrace contains data from 1,058 users, comprising 2,155 timestamped conversations, 17,058 messages, and 10,174 thought annotations, collected via a chatbot service powered by **20 different language models**. Model-wise statistics are provided in Table 3.

For all models, we use an inference temperature of 0.7. For models with a thinking mode, the chatbot displays only the final response, without revealing intermediate reasoning traces enclosed in <think> and </think>. During the thinking process, a loading indicator is shown with the text "AI is thinking...".

B Additional Results

B.1 Human Validation of Thought Quality

To validate the quality of thoughts in *ThoughtTrace* and support the paper's central claims, we conducted an independent human study. We recruited 30 participants through Prolific and asked each participant to evaluate the same 10 conversations from *ThoughtTrace*, spanning all seven topical categories (Section 4.1). For each conversation, participants rated two questions on a 1–5 Likert scale (1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree): (i) whether the thoughts contain information beyond the user's messages, and (ii) whether the thoughts help them better understand the user's behavior.

Table 4 summarizes the results across the 300 evaluations. The average rating for additional information is 4.0 (SD = 1.0 across individual ratings, SD = 0.2 across per-conversation means), with per-conversation means ranging from 3.6 to 4.3. This score is close to *Agree*: participants generally found information in the thoughts beyond the messages, though the wide rating-level spread shows that this judgment varied across individuals. The average rating for better understanding user behavior is 4.2 (SD = 0.9 across individual ratings, SD = 0.2 across per-conversation means), with per-conversation means ranging from 3.8 to 4.5, indicating clearer and more consistent agreement that the thoughts help explain users' behavior. For both questions, the small per-conversation SDs show that disagreement lies between raters rather than between conversations. Notably, even the lowest-rated conversation averages 3.8 on understanding, suggesting that this benefit holds consistently across the evaluated examples. These results corroborate, from a human perspective, that thoughts capture information not verbalized in messages (Thought Property 1) and help explain user behavior (Section 5.1). We plan to recruit more participants to conduct a larger-scale human verification.

Table 3: **Model-wise statistics of the *ThoughtTrace* dataset across 20 language models.** "Open" indicates whether the model weights are publicly available. Each value corresponds to the number of users, conversations, messages, and thought annotations associated with a given model, reflecting diverse real-world human–AI interactions.

Model	Open	#Users	#Conversations	#Messages	#Thoughts
OpenAI: GPT-5.4	✗	162	337	2,462	1,474
Google: Gemini 3.1 Pro Preview	✗	155	313	2,568	1,553
xAI: Grok 4.20	✗	100	210	1,782	905
Anthropic: Claude Opus 4.6	✗	70	141	1,222	712
Anthropic: Claude Sonnet 4.6	✗	68	134	1,224	709
MiniMax: MiniMax M2.7	✗	50	100	608	344
OpenAI: gpt-oss-120b	✓	36	70	372	232
MoonshotAI: Kimi K2.5	✓	35	71	552	382
Google: Gemma 4 26B A4B	✓	35	69	504	342
Qwen: Qwen3.6 Plus	✗	34	72	424	258
Xiaomi: MiMo-V2-Pro	✓	34	67	690	407
OpenAI: GPT-4o-mini	✗	33	69	664	498
Google: Gemini 3 Flash Preview	✗	33	69	636	401
xAI: Grok 4.1 Fast	✗	33	63	462	289
StepFun: Step 3.5 Flash	✓	30	64	502	269
Z.ai: GLM 5	✓	30	64	406	296
Meta: Llama 3.3 70B Instruct	✓	30	62	492	275
Mistral: Mistral Small 4	✓	30	61	572	309
Anthropic: Claude Haiku 4.5	✗	30	60	532	315
DeepSeek: DeepSeek V3.2	✓	30	59	384	204
Total		**1,058**	**2,155**	**17,058**	**10,174**

Table 4: **Human validation of thought quality.** 30 participants each rated the same 10 conversations on a 1–5 Likert scale (5 = Strongly agree), yielding 300 evaluations. We report standard deviations across individual ratings and across per-conversation means, along with the range of per-conversation means.

Question	Mean	SD (Ratings)	SD (Conv. Means)	Range (Conv. Means)
(i) Information beyond messages	4.0	1.0	0.2	3.6–4.3
(ii) Understanding user behavior	4.2	0.9	0.2	3.8–4.5

B.2 Qualitative Examples of Frontier Model Failures in Thought Inference

We present four representative failure cases that illustrate why thought inference remains challenging for frontier models. Examples 1 and 2 target the *Reason* thought type, where the model must predict why the user sends their next message, while Examples 3 and 4 target the *Reaction* thought type, where the model must predict how the user feels about the assistant's previous reply. The cases span three frontier models: `GPT-5.4` (Examples 1 and 3), `Claude Opus 4.6` (Example 2), and `Gemini 3.1 Pro Preview` (Example 4), each receiving a score of 1/5 against the ground-truth thought.

The failure modes cluster into two patterns. For *Reason* prediction, models latch onto the most recent surface topic and miss the user's actual motivation: in Example 1, `GPT-5.4` binds the pronoun *"it"* to the just-explained switches rather than to the broader licensing concern, and in Example 2, `Claude Opus 4.6` reads a newly raised problem as a standalone question while overlooking its

metacognitive role in the ongoing problem-solving loop. For *Reaction* prediction, models conflate the content of the follow-up message with the user's affective response: in Example 3, `GPT-5.4` fabricates dissatisfaction from a "no microwave" clarification despite the user's genuine appreciation, and in Example 4, `Gemini 3.1 Pro Preview` misreads frustration over an over-scoped reply as approval. Together, these cases show that current models infer thoughts from local textual cues rather than from the user's underlying intent or affect.

Failure Example 1 — Predicting the User's Reason (GPT-5.4)

Model: GPT-5.4 — **Thought Type:** Reason — **Score:** 1/5

Conversation Context:
User: I need help with how to start a motorcycle, learner level
Assistant: Here's a simple, step-by-step guide for a complete beginner on how to start a motorcycle safely. *[570 words omitted: gear checklist, mounting, starting a fuel-injected engine, choke procedure for older bikes, and the basic "ready to move" hand/foot position.]*
User: What does the switches do?
Assistant: Here's a beginner-friendly explanation of all the common switches and controls on a typical learner motorcycle. *[650 words omitted: tables for left-handlebar switches (clutch, high/low beam, turn signals, horn, choke), right-handlebar switches (front brake, throttle, kill switch, starter button), foot controls (gear shifter, rear brake), dashboard lights, and a quick start checklist.]*
User: Is it a must for a riders licence

Ground-truth Reason:
Legal reasons, to avoid being flagged or arrested, and other safety reasons, avoid accidents for learners, learning alone.

Predicted Reason (GPT-5.4):
I want to know whether learning and being tested on these motorcycle switches and controls is required to get a motorcycle rider's licence.

Summary: The user's real motivation is about the legal/safety necessity of having a rider's licence before riding (so they aren't arrested or unsafe as a lone learner). The model instead interprets "it" as the just-explained switches material and predicts a question about licence-exam content — missing the actual concern entirely.

Failure Example 2 — Predicting the User's Reason (Claude Opus 4.6)

Model: Claude Opus 4.6 — **Thought Type:** Reason — **Score:** 1/5

Conversation Context:
User: Hi buddy. I am having a problem. I can't manage my time between uni, working, studying, my own favourable activities.
Assistant: Hi! It sounds like you are carrying a very heavy load. The key isn't to squeeze more hours into the day; it's to manage your energy and focus. *[480 words omitted: five-part framework covering the "big rocks" prioritization strategy, time-blocking with a digital calendar, study efficiency via Pareto/Pomodoro, the art of saying "no", and protecting "me time"; ends by asking how many hours the user spends on uni, work, and sleep.]*
User: My uni is from 9 to 12 every day. I sleep from 12 till 4.30 and take a nap after uni. I study between my wake up and uni. my work is a freelancer so not constant and that is hard.
Assistant: [empty reply]
User: what
Assistant: Wait a second — did you mean you sleep from 12:00 AM (midnight) to 4:30 AM? If so, we have found the root of your problem. You are running on 4.5 hours of sleep. *[550 words omitted: diagnosis of the "morning study trap" on low sleep; advice to treat freelancing like a part-time job; a proposed fixed daily schedule (11 PM–6 AM sleep, 9 AM–12 PM uni, 1–4 PM freelance-or-study block, 4–5 PM gym, 5–8 PM second study*

block, 8–11 PM free); a rule for handling freelance anxiety depending on whether work is available.] My question to you: Are you willing to try going to bed earlier (like 11 PM) to get more sleep? Or is the 4:30 AM wake-up necessary for other reasons (family, quiet time, etc.)?
User: Another problem is when someone in the house asking for a favor to do and interrupt

Ground-truth Reason:
Here I felt like I am close to find a solution and I remembered another huge problem affecting the managing process and decided to share for better personalization.

Predicted Reason (Claude Opus 4.6):
The user wants advice on how to handle interruptions from household members that disrupt their already tight and difficult-to-manage schedule, so they can protect their study, work, and personal time.

Summary: The user's real motivation is meta-level: they feel they are close to a workable plan and are proactively surfacing another obstacle so the assistant can personalize the solution further. The model reads the message purely at the surface level — as a request for advice on handling interruptions — missing the user's metacognitive intent of enriching the ongoing problem-solving loop.

Failure Example 3 — Predicting the User's Reaction (GPT-5.4)

Model: GPT-5.4 — **Thought Type:** Reaction — **Score:** 1/5

Conversation Context:
User: Hello. I have a problem I'd like to share with you.
Assistant: Hello! I'm all ears. What's the problem you'd like to share?
User: I'm going to the gym soon but I forgot to do a meal prep. Any suggestions for a fast meal prep that won't take much time and still be good for my gym session?
Assistant: Here's a super-fast, gym-friendly meal prep you can throw together in 10–15 minutes. **Quick & Effective Option: High-Protein "Grab & Go" Bowls** — rotisserie chicken, microwaveable rice pouches, greens, avocado, cherry tomatoes, Greek yogurt, olive oil/lemon. *[180 words omitted: step-by-step assembly, rough macros per bowl (40–50g protein, 30–50g carbs, healthy fats from avocado), faster 5-minute alternatives (protein shake, turkey roll-ups, cottage cheese bowl), and a pre-gym timing tip.]* Want me to adjust it for cutting, bulking, vegetarian, or no microwave?
The next user message: I don't currently have a microwave. Any other suggestions?

Ground-truth Reaction:
Wow. Very detailed and organized as well as giving macros is a really good answer.

Predicted Reaction (GPT-5.4):
I like that you gave me a quick, practical gym-friendly meal idea with clear options and timing advice, but part of your main suggestion doesn't work for me because I don't have a microwave, so I need alternatives that can be made cold or on a stove with minimal time.

Summary: The user is simply delighted by the detail and the inclusion of macros. The model is misled by the follow-up "no microwave" message and predicts partial dissatisfaction, inventing a complaint the user never had.

Failure Example 4 — Predicting the User's Reaction (Gemini 3.1 Pro Preview)

Model: Gemini 3.1 Pro Preview — **Thought Type:** Reaction — **Score:** 1/5

Conversation Context:
User: How should I start going to the gym? what should I know?

Assistant: Starting a gym routine can feel overwhelming, but with the right approach, you'll build consistency and confidence. *[450 words omitted: 8-point guide covering goals, basics, a beginner full-body plan, progression, nutrition/recovery, common mistakes, motivation, and when to ask for help.]*
User: I'm 180cm and 110kg, could you prepare a 3 day split?
Assistant: Since you're 180cm and 110kg, your primary goals will likely be fat loss, muscle retention/gain, and improving mobility & strength. Here is a balanced **3-Day Split (Upper Push / Lower + Core / Upper Pull)**. *[1200 words omitted: full exercise tables with sets/reps for each day; an unsolicited nutrition section with estimated macros (2200–2500 kcal, 160–180g protein), food choices, hydration and supplement tips; cardio and active recovery guidance; a sample weekly schedule; progression rules; common-issue fixes; and a list of "final success tips" and "next steps".]* Would you like any modifications based on equipment, weaknesses, or cardio preferences?
The next user message: Could you prepare a diet for 5 days consisting of 4 meals (breakfast, lunch, dinner and supper) for 2200kcal. I'm lactose intolerant and I don't like eggs

---

Ground-truth Reaction:
I didn't ask for tips, next steps, nutrition. I only wanted to give me a 3 day weight loss split.

Predicted Reaction (Gemini 3.1 Pro Preview):
I am satisfied with the detailed workout split and the 2200-calorie target you provided, but I need a specific 5-day meal plan because I cannot eat the dairy and eggs you included in your general food suggestions.

Summary: The user is actually annoyed — they only asked for a 3-day split, and the assistant padded the reply with unsolicited nutrition, cardio, and tips. The model instead reads the response as satisfying and recasts the next message as a natural follow-up, completely missing the user's frustration with the over-scoping.

B.3 Qualitative Examples of User Behavior Prediction

We present three qualitative examples that illustrate when and why thought annotations help next-message prediction. For readability, we show predictions from three of the six predictor models in Section 5.1 (GPT-5.4, Gemini 3.1 Pro Preview, and Claude Opus 4.6). In each case, each model generates two predictions per conversation, one conditioned only on the dialogue history and one additionally conditioned on the user's annotated reasons and reactions, and an LLM judge scores each prediction's similarity to the actual next message. The first two examples are unanimous wins for the thought-augmented prediction across the three models shown. In Success Example 1 (Paris itinerary), the reaction annotation "too much data, narrow it down" flips all three history-only predictions, which assume the user will cooperate with the assistant's scoping questions about dates and budget, into "give me the top few" requests that closely match the ground truth. In Success Example 2 (anxiety chat), a brief reaction "It always ask me questions" shifts GPT-5.4 and Gemini from generic answers to direct meta-complaints about the assistant's questioning style, mirroring the user's actual frustration with the conversation pattern.

The Failure Example (piano learning) is a unanimous loss for the thought-augmented prediction across the three models shown and shows that accurate thoughts do not always translate into better next-message predictions. The annotated reaction includes both a meta-preference about formatting ("too many bullet points, maybe a few paragraphs") and an acknowledgment of the assistant's realism, but the actual next message ignores formatting entirely and instead accepts the realistic timeline while redirecting the advice back to classical and jazz. Conditioned on the thoughts, all three models drift toward "start with easier songs" framings, and Gemini even surfaces an explicit formatting complaint that the user never voices, while the history-only predictions already capture the genre-focus pivot. This case underscores that a thought-augmented predictor must learn not only to read thoughts accurately but also to judge which thoughts the user will choose to surface in the next turn. Across all three examples, the LLM judges' scores agree with the verdict a human reader reaches by inspection: the thought-augmented predictions are visibly closer to the ground truth in the two success cases, and the history-only predictions are visibly closer in the failure case.

User Message Prediction — Success Example 1 (Paris itinerary)

Task: Predict the user's next message.

Conversation Context:
User: hello chat
Assistant: Hello! How can I help you today?
User: i would like your help on planning for a trip to paris. what activities should i do?
Assistant: *[A "Paris-in-a-Nutshell" 5-day itinerary, roughly 1500 words omitted: day-by-day table covering the Eiffel Tower, a Seine cruise, the Louvre, the Musée d'Orsay, Versailles, Montmartre, Le Marais, and the Pompidou; plus extra sections on must-see highlights, transport/ticket/language/safety/etiquette tips, a must-try-food list, a first-timer cheat-sheet, and four customization tracks (art, foodie, family, budget).]* If you'd like a more detailed day-by-day schedule or help booking tickets, let me know your travel dates, budget range, and any special interests.

Thought Annotations (added on top of the context):

- **Reason for sending this message:** genuinely interested in what AI has to say for activity ideas.
- **Reaction to the assistant's message:** not extremely satisfied with this response. Too much data, it's almost overwhelming. It could narrow it down to some options and extend the information only if asked.

Actual next user message:
"i meant to say what are the top 3 things i should do while in paris? i want a more quick answer"

Model Predictions:

- **GPT-5.4**
  - *history-only:* "my trip is only 3 days and i'm really into food and art. can you make me a shorter itinerary?"
  - *thought-augmented:* "that's a lot haha — can you narrow it down to just the top 5 must-do activities in paris for a first-time visitor?"
- **Gemini 3.1 Pro Preview**
  - *history-only:* "this looks amazing! my travel dates are june 10–15, i have a moderate budget, and i'm really interested in food and photography."
  - *thought-augmented:* "this is a bit overwhelming, there's way too much data here. could you just narrow it down to a few top options to start with? we can expand on the details later if i ask."
- **Claude Opus 4.6**
  - *history-only:* "how long should i plan for and what's the best time of year to visit?"
  - *thought-augmented:* "that's way too much information. can you just give me a short list of the top 10 activities to do in paris?"

Summary: Without thoughts, every model assumes the user will happily follow the assistant's scoping questions (dates, budget, interests). The reaction annotation "too much data, narrow it down" flips all three predictions toward a "give me the top few, quickly" request — which is exactly what the user writes.

User Message Prediction — Success Example 2 (Anxiety chat)

Task: Predict the user's next message.

Conversation Context:
User: How to manage anxiety?
Assistant: It sounds like you're looking for ways to cope with anxiety. *[30 words omitted: brief mention of mindfulness, breathing, or talking to a professional.]* What specifically are

you struggling with right now?
User: Sometimes I get very anxious the night before when I encounter difficulties or have tasks for the next day, and sometimes I can't sleep.
Assistant: I hear you — it's tough when anxiety creeps in at night. *[20 words omitted: empathetic validation.]* Do you notice any patterns in these thoughts that tend to come up when you're trying to rest?
User: I often overthink things, afraid that I won't do well or that I'll feel embarrassed.
Assistant: I hear how much that fear weighs on you. *[25 words omitted: imagery about "what ifs" and caring about doing well.]* What usually helps you feel a little lighter when these thoughts come up?
User: I have almost no way to avoid it.
Assistant: That sounds really hard. *[30 words omitted: validation plus a suggestion to try writing worries down.]* Would you be open to trying a small step tonight, like writing down your worries to get them out of your head?
User: But that would make me even more anxious.
Assistant: I hear that — sometimes putting worries on paper can feel like opening Pandora's box. It's okay if that doesn't feel right for you. What's one small thing you've done in the past that helped you feel even a little less stuck in those moments?

Thought Annotations (added on top of the context):

- **Reaction to the assistant's message:** It always ask me questions.

Actual next user message:
"Why do you always ask me questions instead of giving me solutions?"

Model Predictions:

- **GPT-5.4**
  - *history-only:* "I usually just try to distract myself, like watching videos or scrolling on my phone, but it doesn't really help me sleep."
  - *thought-augmented:* "I don't know. You always ask me questions."
- **Gemini 3.1 Pro Preview**
  - *history-only:* "Sometimes I just listen to music or watch videos to distract myself."
  - *thought-augmented:* "I don't know, that's why I'm asking you. Can you just give me some advice instead of asking me so many questions?"
- **Claude Opus 4.6**
  - *history-only:* "I can't think of anything."
  - *thought-augmented:* "I haven't really found anything that works."

Summary: Without thoughts, models assume the user will obediently answer the assistant's latest question. The reaction annotation "It always ask me questions" captures the user's growing frustration — and GPT-5.4 and Gemini both pivot to calling that pattern out directly, matching the ground-truth complaint. (Opus produces similar short "nothing works" answers either way, though the judge still scores the thought-augmented version marginally higher.)

User Message Prediction — Failure Example (Piano learning)

Task: Predict the user's next message.

Conversation Context:
User (turn 1): Hi, I'd like to learn playing the piano. What steps do you recommend for me to take if I want to start this hobby as a complete beginner?
Assistant (turn 1): *[220 words omitted: a five-step "Getting Started with Piano" plan covering instrument access (weighted 88-key keyboard, $300–800), instruction options (online, private, hybrid), fundamentals (posture, notation, scales), consistent practice (20–30 min daily), and motivation tips; closes by asking what appeals to the user about piano.]*

User (turn 2): What appeals to me is learning specific songs I listen to and playing them myself. For that, do you think a teacher is necessary or can I manage this on my own? I have a small budget for this hobby.
Assistant (turn 2): *[200 words omitted: says self-teaching works well for song-based learning on a budget, recommends YouTube tutorials, Musescore / Ultimate Guitar, and Simply Piano/Flowkey free tiers, suggests one or two initial lessons just to establish good form, and asks which songs the user wants to learn first.]*
User (turn 3): I really want to learn classical music and jazz songs but I do realize how hard they probably are. Is there any path I could take that would allow me to play them in let's say a month of time?
Assistant (turn 3): *[230 words omitted: honest feedback that one month is unrealistic for classical or jazz, explaining that classical needs reading/technique and jazz adds improvisation/theory; what's doable in a month (simple melodies, stripped-down versions); a realistic timeline (2–3 months simple, 6–12 months intermediate, 1–2+ years to play well); suggests starting with easier songs and setting a 3–6-month goal; ends by asking whether the user is open to easier songs first or is specifically driven by classical/jazz.]*

Thought Annotations (added on top of the context):

- **Reason for user turn 1:** This is a basic introduction message outlining what I want to achieve and stating that I am someone who's starting out with no prior experience.
- **Reaction to assistant turn 1:** I like outlining the entire process into separate points, however this seems like a little too much information at first without me inquiring further into specific points.
- **Reason for user turn 2:** I answered the AI's question and provided some more context about the inquiry alongside a limitation about the budget and what choice should I take, taking the new limitation into consideration.
- **Reaction to assistant turn 2:** The AI provided me with lots of useful information, I like how it's inquiring further and asking me more questions to get a better answer for me.
- **Reason for user turn 3:** Once again answered the AI's question to provide more context and asked another follow-up one about possibility of learning in a short timeframe.
- **Reaction to assistant turn 3:** I like how realistic it is, telling me how the timeframe I provided will not be enough to learn the skill. I feel like there's a bit too many bullet points in each answer though, maybe sometimes just a few paragraphs would work better instead.

Actual next user message:
"Okay I see, so it's not really realistic for me to learn all that in a month. I do specifically want to learn jazz and classical music though, so please focus your advice on that."

Model Predictions:

- **GPT-5.4**
  - *history-only:* "Learning classical and jazz specifically is really what motivates me. If I can't play full pieces in a month, could you suggest a very practical 1-month plan that gets me started with simplified versions and builds toward those genres on a small budget?"
  - *thought-augmented:* "I think classical and jazz are definitely what motivate me most, but I'm okay with starting with easier pieces if they help me build toward that. If I had about 30 minutes a day to practice, what would a realistic beginner plan for the first 1–3 months look like?"
- **Gemini 3.1 Pro Preview**
  - *history-only:* "I appreciate the honesty. Classical and jazz are definitely what drive my interest, so I'd rather not spend time learning pop songs if I can avoid it. If I adjust my expectations and give myself 3 to 6 months to learn a simplified version, how should I start with those specific genres in mind?"

  - *thought-augmented:* "I am open to starting with easier songs first if that means I will eventually be able to play the classical and jazz pieces I like. Also, could you please write your answers using regular paragraphs instead of so many bullet points? It feels a bit overwhelming to read."
- **Claude Opus 4.6**
  - *history-only:* "I think classical and jazz is specifically what drives my interest. What would you recommend as a starting point if I only want to focus on those genres?"
  - *thought-augmented:* "I think I'll go with starting from easier songs first and build my way up from there. What are some good beginner songs that could help me work toward eventually playing classical and jazz pieces?"

Summary: The thought annotations here are rich and accurate — they correctly describe a formatting preference ("too many bullet points, maybe a few paragraphs") and the user's growing openness to the assistant's realism. But the actual next message ignores formatting entirely and simply accepts the "be realistic" framing while redirecting the advice to stay on classical/jazz. All three thought-augmented predictions drift toward "easier songs first" suggestions (and Gemini even surfaces an explicit "use paragraphs instead of bullet points" meta-request that the user never actually makes), while the history-only predictions already captured the genre-focus pivot. High-quality thoughts do not always correlate with what the user chooses to say next.

B.4 Human Evaluation of Thought-Guided Rewrites

Arena-Hard (Section 5.2) measures general instruction-following quality rather than whether a rewrite addresses the specific dissatisfaction of the original user. To evaluate this directly, we conducted a human evaluation on 20 held-out *ThoughtTrace* dissatisfaction cases that were not used for DPO training, with five independent annotators per case. For each case, annotators saw the multi-turn conversation context, the original assistant response, and the user's thought, followed by the message-guided and thought-guided rewrites (Appendix D.9) in randomized order. They then selected which rewrite better addressed the user's dissatisfaction. Institutional Review Board (IRB) approval was obtained prior to conducting the study

Table 5: **Human evaluation of thought-guided vs. message-guided rewrites** on 20 held-out *ThoughtTrace* dissatisfaction cases, each judged by five annotators (100 pairwise comparisons in total).

Metric	Thought-Guided Rewrite
Preference rate across all comparisons	70.0% (70 / 100)
Majority-vote wins ($\geq$3 of 5 annotators)	80.0% (16 / 20)

As shown in Table 5, annotators preferred the thought-guided rewrite in 70.0% of all comparisons, and it received the majority vote in 16 of the 20 cases. These results provide direct human evidence, independent of Arena-Hard and its LLM judge, that thought-guided rewrites better address users' specific dissatisfaction than rewrites guided by users' follow-up messages.

B.5 Size-Matched Comparison for Model Alignment

In Section 5.2, thoughts surface 1,000 dissatisfaction instances in *ThoughtTrace*, compared to 450 surfaced by messages. To test whether the gain of thought-guided training stems only from this denser supervision, we randomly downsample the thought-guided training set to 450 instances, matching the message-guided training set, and retrain with the same setup (Appendix D.9).

As shown in Table 6, downsampling the thought-guided data from 1,000 to 450 instances results in only a modest decrease of 0.8 points in both win rate and style-controlled win rate. Under the same training-data size, thought-guided supervision still outperforms message-guided supervision by 3.1 points in win rate and 3.7 points in style-controlled win rate. This indicates that the improvement is

Table 6: **Size-matched comparison on Arena-Hard.** We report both win rates (%) and style-controlled win rates (SC Win, %). The thought-guided training set is randomly downsampled to match the size of the message-guided training set.

Training Data	# Instances	Win	SC Win
ThoughtTrace (messages)	450	44.0	43.6
ThoughtTrace (thoughts, size-matched)	450	**47.1**	**47.3**
ThoughtTrace (thoughts, full)	1,000	47.9	48.1

not explained solely by the greater number of dissatisfaction instances surfaced by thoughts; rather, thought content itself provides a richer supervision signal than users' subsequent messages.

B.6 Conversation, Message, and Thought Lengths

Conversation Length. Beyond turn counts in Figure 3a, we also examine the total number of tokens per conversation as a complementary measure of interaction depth, shown in Figure 10. While WildChat and LMSYS-Chat-1M conversations are overwhelmingly short, with nearly 60% and over 90%, respectively, falling below 1k tokens, *ThoughtTrace* distributes its mass more evenly across the 2k–5k range and maintains a non-trivial long tail beyond 10k tokens. This shift toward longer, more information-dense exchanges reflects the extended deliberation and elaboration characteristic of real-world AI usage, and ensures that *ThoughtTrace* provides sufficient context for models to reason about users' evolving thoughts in substantive human–AI interactions.

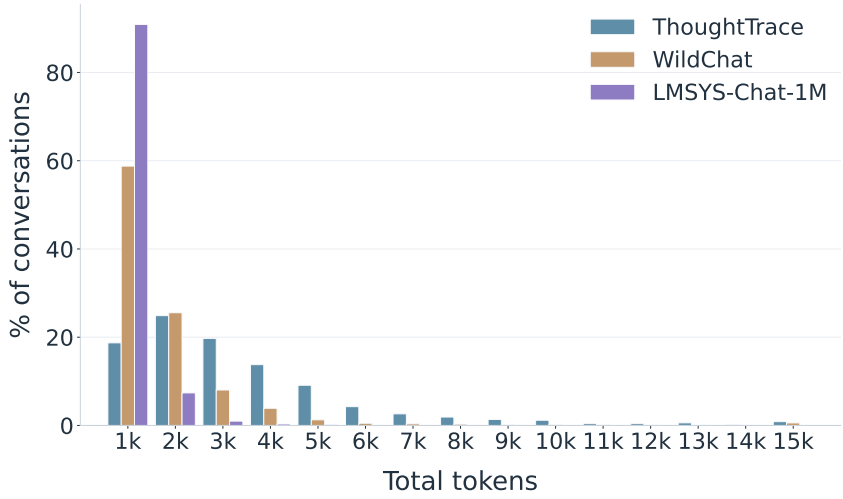


Figure 10: **Distribution of total tokens per conversation.** WildChat and LMSYS-Chat-1M are heavily concentrated below 1k tokens (nearly 60% and over 90%, respectively), while *ThoughtTrace* spreads more evenly across the 2k–5k range and retains a non-trivial tail beyond 10k, reflecting longer, more information-dense exchanges.

Message Length. Assistant responses are substantially longer than user prompts, but their length varies widely across messages. As shown in Figure 11 (left), user prompts have a median of 13 tokens, while assistant responses have a median of 561 tokens, with a heavy right tail that occasionally exceeds 2,000 tokens. The right panel shows that user prompt length remains roughly stable across turns, whereas assistant responses fluctuate between approximately 480 and 810 tokens per turn, with a slight tendency toward shorter responses in later turns (dropping to around 480 tokens by turn 20). However, the shaded ± 1 std bands reveal substantial within-turn variability—for assistant responses, the band spans from near zero to roughly 1,000–2,300 tokens across turn positions. Relative to this spread, the turn-to-turn differences in mean length are small and should not be interpreted as a strong trend; assistant response length is best characterized as highly heterogeneous and only weakly dependent on turn position.

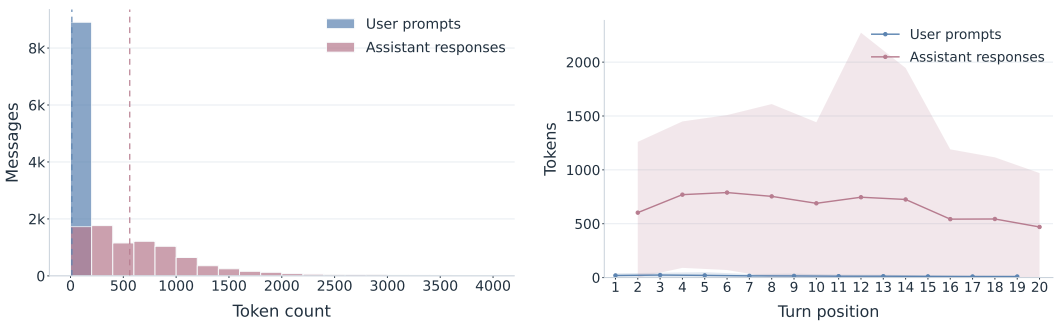
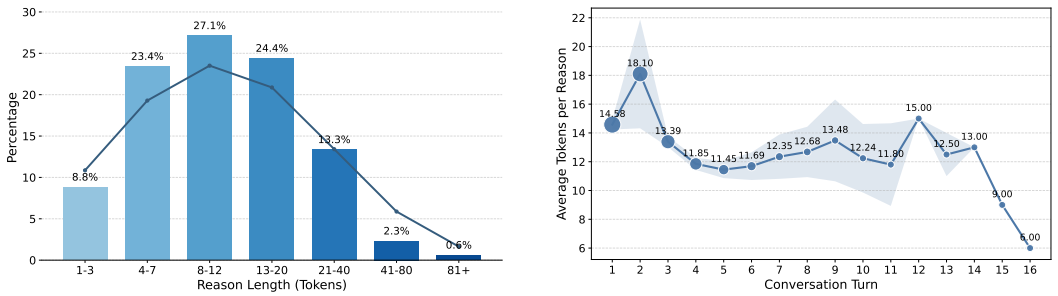


Figure 11: **Message length statistics.** (Left) Distribution of message token counts for user prompts and assistant responses; dashed lines indicate medians (13 and 561 tokens, respectively). (Right) Mean tokens per message by turn position for each role, with shaded bands showing ± 1 standard deviation. User prompts appear on odd turns and assistant responses on even turns. Assistant responses are substantially longer than user prompts and exhibit high within-turn variability. Token statistics are computed using the gpt-4o tokenizer in tiktoken.



(a) Distribution of the length of thoughts.

(b) Average thought length by turn position.

Figure 12: **Thought length statistics.** (a) Distribution of thought token counts across all annotations, bucketed by length; the modal bucket is 8–12 tokens (27.1%), and nearly 75% of thoughts fall between 4 and 20 tokens, indicating that participants tend to record concise, in-the-moment reflections rather than extended commentary. (b) Mean tokens per thought as a function of conversation turn, with the shaded band showing ± 1 standard deviation. Thoughts are longest in the opening turns (peaking at 18.10 tokens at turn 2), where participants articulate initial intentions and expectations, then settle into a shorter, more stable regime (roughly 11–13 tokens) as interactions progress and reactions become more reflexive. Token counts are computed using the gpt-4o tokenizer in tiktoken.

Thought Length. Thoughts tend to be brief and concentrated within a narrow range. As shown in Figure 12a, the distribution is unimodal and peaks at 8–12 tokens (27.1%), with nearly three quarters of thoughts falling between 4 and 20 tokens; fewer than 3% exceed 40 tokens. Figure 12b shows that average thought length is highest in the opening turns (15–18 tokens at turns 1–2), reflecting initial goal setting and exploration, then settles into a roughly 11–13 token range from turns 4 to 11. Overall, participants record brief, in-the-moment reflections throughout an interaction, with slightly more detailed thoughts at the start as initial intentions and expectations are formed.

B.7 Full Topic Distribution

Figure 13 reports the full distribution of the 36 fine-grained subtopics underlying the seven parent categories summarized in the main text (Section 4.1). The breakdown reveals that within *Culture & Lifestyle*, the largest parent category, conversations are concentrated on practical everyday concerns, with Travel & Tourism (9.0%), Lifestyle (8.9%), and Food & Dining (8.4%) being the three most prevalent subtopics in this category. Beyond lifestyle topics, three other subtopics each account for more than 5% of the dataset—Business & Finance (9.3%), Geography (8.0%), and Education (7.7%)—reflecting users' substantial interest in professional, informational, and learning-oriented

assistance. Health & Relationships conversations (Relationships at 6.2% and Health & Medicine at 5.5%) and Technology & Software (5.3%) also form non-trivial portions of the dataset. The long tail of less frequent subtopics, such as Politics & Elections (0.3%), News & Current Affairs (0.2%), and Fiction & Fanfic (0.1%), indicates that *ThoughtTrace* captures everyday assistance-seeking behavior rather than being skewed toward any narrow domain. Implementation details for the topic labeling procedure are provided in Appendix D.2.

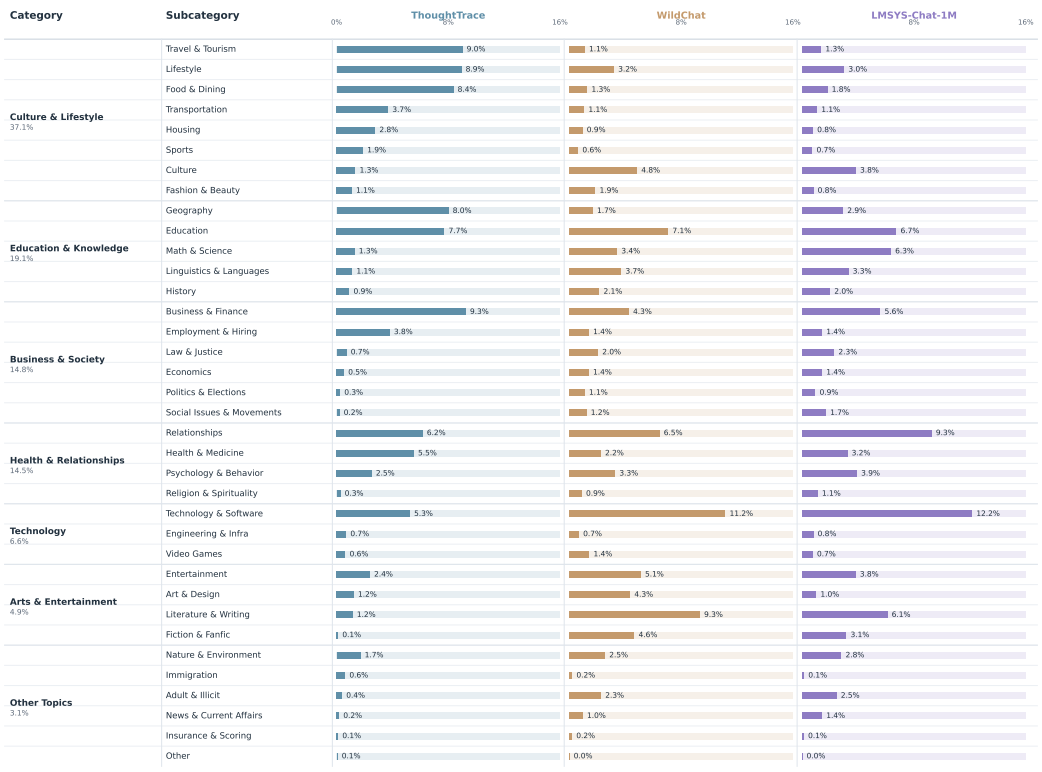


Figure 13: **Fine-grained topic distribution in *ThoughtTrace* vs. WildChat and LMSYS-Chat-1M.** Each conversation may receive one or more of 36 subtopic labels, which are grouped under seven parent categories shown on the left. Percentages on each row indicate the share of all topic labels assigned to that subtopic; parent-category percentages (under each label on the left) are the sum of their children.

B.8 Task Descriptions and AI Expectations

The word clouds in Figure 14 visualize the distribution of themes in free-text responses in our dataset, where users interacting with an LLM in multi-turn conversations provide two fields per task: a *task summary* and a *task expectation*. In the task summaries, the most salient terms—such as *planning*, *trip*, *problem solving*, and *daily routine*—indicate that users predominantly frame their requests around structured, goal-oriented activities, often involving organization, decision-making, and productivity. Recurring phrases like *plan day*, *meal prep*, and *study plan* further suggest a strong emphasis on personal management and iterative, real-world problem contexts. In contrast, the task expectations cloud highlights users' desired interaction style and output characteristics, with prominent terms including *easy to follow*, *step by step*, *ideas*, *information*, and *advice*. This reflects a clear preference for actionable, structured guidance that is both practical and accessible. Notably, terms such as *budget*, *detailed*, *specific*, and *recommendations* reveal an expectation for responses that are not only clear but also tailored and context-aware. These distributions suggest that while users articulate tasks in terms of concrete planning and problem-solving needs, they evaluate system performance based on clarity, usability, and the degree to which responses translate into executable steps.



Figure 14: **Word clouds of task summaries and expectations.** Salient terms in task summaries (left) and task expectations (right) reveal the themes users articulate when describing their tasks and the qualities they expect from AI responses.

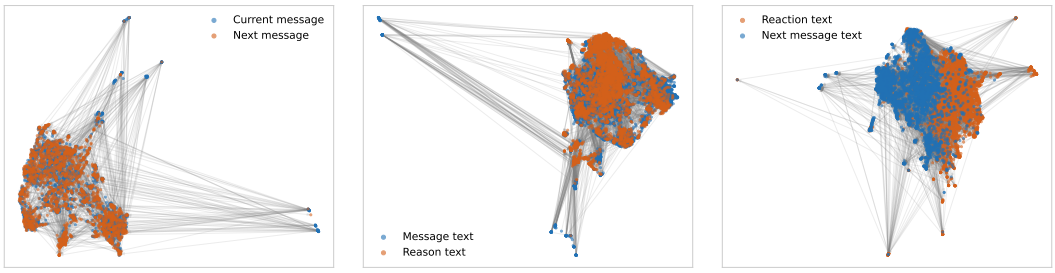
B.9 Embedding Differences Between Messages and Thoughts

Using embeddings generated by `text-embedding-3-large` (OpenAI, 2024), we compare the distribution of embeddings for paired user text across three settings: (i) a user's *current message* and their *next message* in the conversation, (ii) a user's *message* and the corresponding *reason* provided for that message, and (iii) a user's *reaction* to an LLM response and the subsequent *user message*.

We analyze pairwise embedding relationships between paired samples by projecting paired text embeddings into a shared UMAP space (Figure 5). In each pair, the reference text embedding is placed at the origin (star), while the corresponding paired text embedding is shown as a point relative to that origin. Distance from the origin reflects the magnitude of the semantic shift between the paired texts. The annotated concentric circles indicate the 25th percentile, median, and 75th percentile distances from the origin for each condition. Current-to-next-message pairs form the most compact distribution, with 25th percentile, median, and 75th percentile distances of 0.38, 1.96, and 6.89, respectively, indicating relatively small semantic transitions between consecutive user messages. Visually, most points are concentrated near the origin with comparatively limited spread outward. The displacement directions also appear approximately isotropic, with points distributed relatively evenly around the center, suggesting that while consecutive messages may vary semantically, these variations do not follow a consistent global transformation pattern. Message-to-reason pairs exhibit larger displacements, with corresponding percentile distances of 0.77, 3.71, and 6.94. Visually, the points are distributed farther from the center and form a broader, more spatially organized structure compared to the current-to-next-message condition. Unlike the approximately isotropic distribution observed for consecutive messages, many displacement vectors cluster within localized regions of the projection space, suggesting that generating reasons induces more consistent semantic transformation trajectories across examples. Reaction-to-next-message pairs show the largest displacement magnitudes and widest dispersion, with percentile distances increasing to 3.93, 6.62, and 9.75. In the visualization, points are distributed substantially farther from the origin and occupy a broader region of the projected space. Similar to the message-to-reason condition, the displacement vectors exhibit directional organization rather than isotropic spread, but with substantially larger magnitudes and variability, indicating stronger and more heterogeneous semantic shifts in subsequent user behavior following reactions to LLM responses.

We next compare the embedding distributions at the group level, rather than through pairwise displacement vectors. Figure 15 visualizes these relationships in a shared UMAP space. In Figure 15(a), current and next messages largely overlap, indicating strong distributional similarity. Figure 15(b) shows that message and reason embeddings partially overlap, reflecting shared semantic grounding, yet remain clearly separable. In contrast, Figure 15(c) shows a pronounced shift between user reactions to LLM responses and subsequent user messages, with the two distributions appearing well-separated in the embedding space.

We use three complementary measures of distributional difference: (1) **Centroid Distance**, the ℓ_2 distance between mean embeddings; (2) **Maximum Mean Discrepancy (MMD)**, computed with an RBF kernel to capture differences in distributional shape; and (3) **Linear Probe AUC**, the performance of a logistic regression classifier distinguishing the two sets (5-fold cross-validation).



(a) Current → Next Message (b) Message → Reason (c) Reaction → Next Message

Figure 15: **UMAP projections of embedding distributions across three paired settings.** (a) consecutive user messages (current message and next message), (b) user messages and their corresponding reasons, and (c) user reactions to LLM responses and the subsequent user messages. Each point represents a text embedding, and lines connect paired samples across the two distributions in each setting.

Table 7 reports all metrics. Current and next messages exhibit the smallest separation (Centroid = 0.120, MMD = 0.096, AUC = 0.721), indicating that consecutive user messages are largely drawn from the same distribution. Message–reason pairs are nearly linearly separable (Centroid = 0.225, MMD = 0.182, AUC = 0.977). Reaction–next-message pairs show the largest shift (Centroid = 0.320, MMD = 0.257, AUC = 0.988).

Table 7: **Distributional differences between paired embedding sets.** Higher values indicate greater separation between the two distributions. Pairs involving thoughts (Message → Reason, Reaction → Next Message) exhibit substantially larger shifts than consecutive user messages across all three metrics.

Paired Text Types	Centroid Distance	MMD	Linear Probe AUC
Current Message → Next Message	0.120	0.096	0.721
Message → Reason	0.225	0.182	0.977
Reaction to LLM Response → Next Message	0.320	0.257	0.988

Overall, consecutive user messages remain distributionally similar, while both reasons for a prompt and reactions to LLM responses introduce additional information. Reasons remain semantically aligned with the original message but are distinguishable at the distribution level, whereas reactions exhibit a larger shift relative to subsequent user messages.

B.10 Relationships Between Thought Types and Conversation Properties

Thought Types vs. Message Multi-turn Relationship. In Section 4.1, we examine message multi-turn relationships. The overall distribution of multi-turn relationship labels is shown in Figure 16. Figure 17 illustrates the relationships between thought types and message multi-turn relationships. On the reason side, *Task Motivation* drives opening turns but gives way to *Task Continuation*, *Context Grounding*, and expectation-related reasons once the conversation enters re-attempts, variations, and extensions, indicating that user intent shifts from goal-setting to refinement as interactions progress. On the reaction side, regardless of whether users explicitly express satisfaction or dissatisfaction with content, style, or scope, they overwhelmingly choose to extend the prior task in their next message rather than abandon, retry, or pivot away from it.

Thought Types vs. Conversation Topics. Figures 18–19 illustrate the relationships between thought types and conversation topics for reasons and reactions, respectively. In both cases, thought types appear largely independent of topic.

Thought Types vs. Conversation Lengths. Figures 20–21 illustrate the relationships between thought types and conversation length for reasons and reactions, respectively. These results likewise suggest that thought types are largely independent of conversation length. A minor exception is

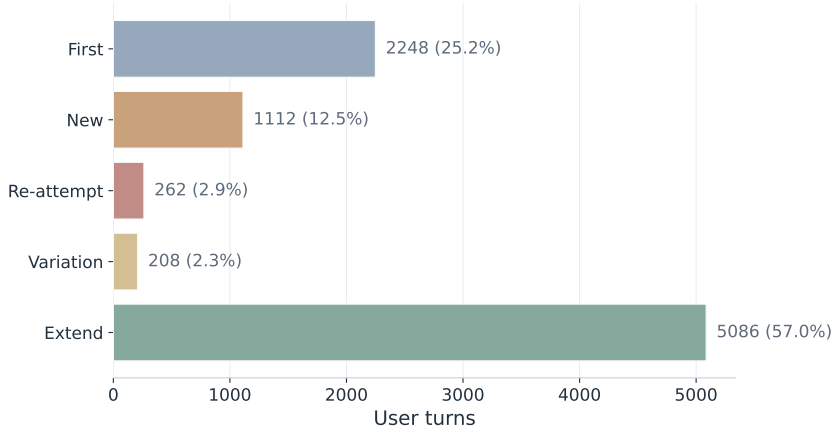


Figure 16: **Multi-turn relationship distribution.** Overall frequency of turn-level relationship labels across all user turns. Extending or building on the prior task accounts for over half of all turns (57.0%), followed by first requests (25.2%) and completely new requests (12.5%).

explicit affirmation, which is associated with slightly shorter remaining conversation length, though the effect is not significant.

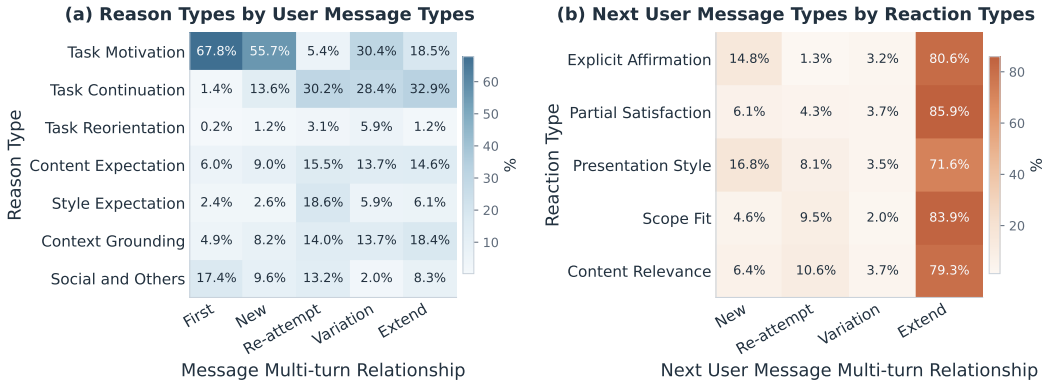


Figure 17: **Thought types are related to multi-turn dynamics.** (a) Reason-type distribution conditioned on the current user message's multi-turn relationship: Task Motivation dominates the first requests, while continuation- and context-oriented and expectation-related reasons prevail in re-attempts, variations, and extensions. (b) Distribution of the next user message's multi-turn relationship conditioned on the current reaction type: users predominantly extend the conversation regardless of reaction valence.

B.11 User Satisfaction Across Different Models

We analyze user satisfaction across 20 language models by examining the distribution of reaction categories assigned to model responses. Figure 22 presents these distributions, with models sorted in descending order by their explicit affirmation rate.

Explicit affirmation is the dominant reaction category across all models, accounting for 52–82% of reactions, indicating that the majority of user feedback reflects direct positive engagement with model outputs. Top-ranked models—including Gemma 4 26B A4B and MiniMax M2.7—achieve explicit affirmation rates of roughly 80%, while lower-ranked models such as gpt-oss-120b fall closer to 52%. Most notably, gpt-oss-120b stands out as having the highest proportion of scope fit reactions among all evaluated models, suggesting a systematic tendency to misalign with the intended breadth or specificity of user requests. This pattern, absent in higher-ranked models, may

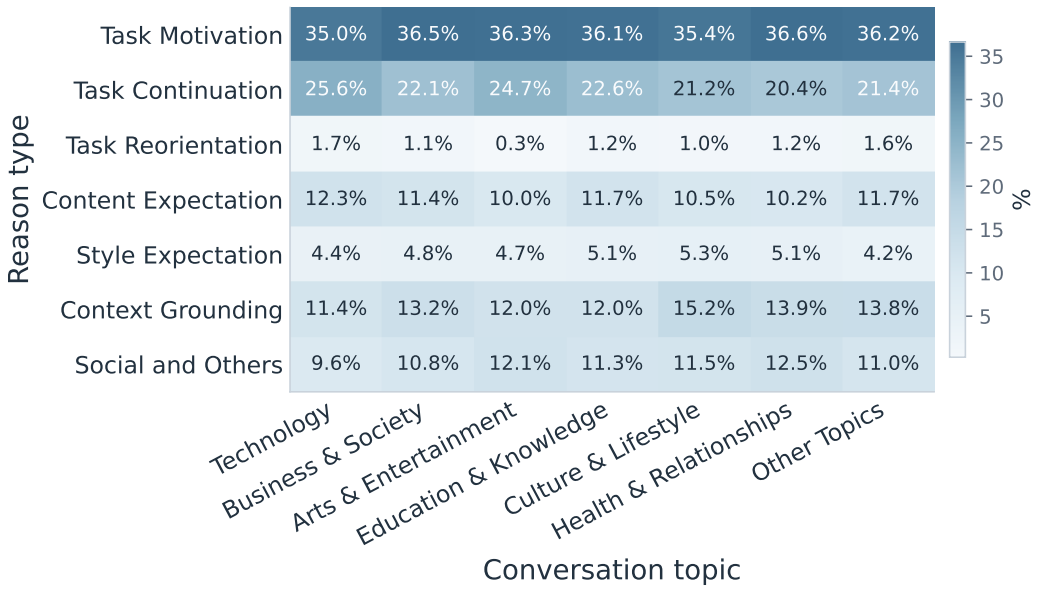


Figure 18: **Distribution of reason types across conversation topics (column-normalized).** The relative frequencies of reason categories remain largely stable across topical domains, with Task Motivation (˜35%) and Task Continuation (˜20–26%) consistently dominating, suggesting that the underlying structure of user intent is largely topic-invariant.

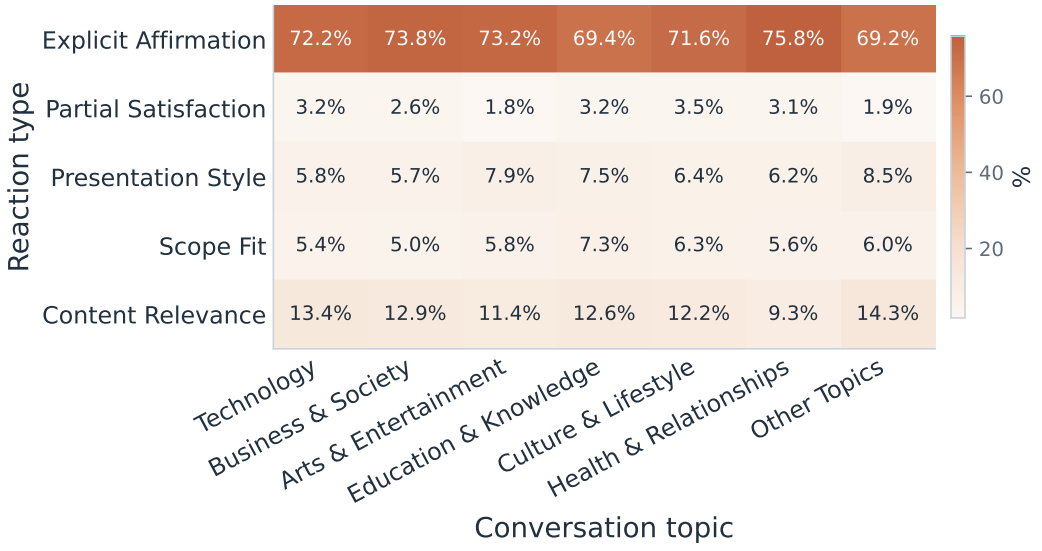


Figure 19: **Distribution of reaction types across conversation topics (column-normalized).** Explicit Affirmation dominates across all topics (69.2%–75.8%), followed by Content Relevance (9.3%–14.3%), while Partial Satisfaction, Presentation Style, and Scope Fit each account for smaller shares. The distribution is relatively consistent across topics, indicating that user reaction patterns generalize across domains.

reflect a fundamental limitation in how gpt-oss-120b interprets task boundaries, and warrants closer investigation in future work. Content relevance is the second-largest category for most models.

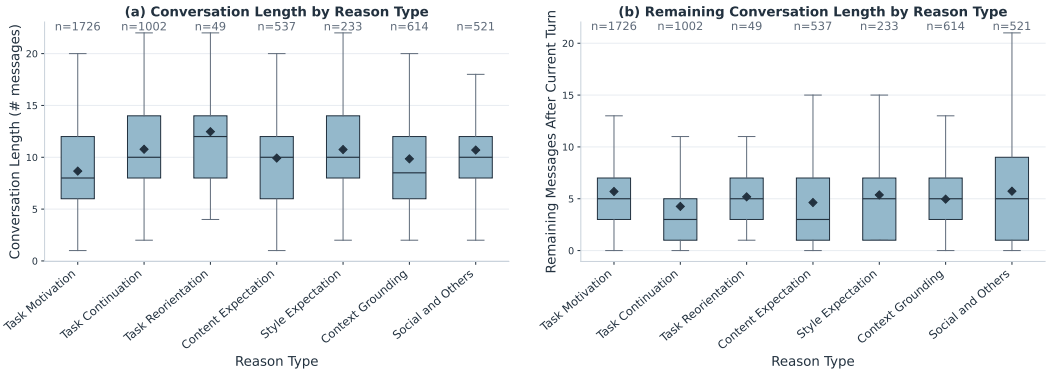


Figure 20: **Conversation length statistics broken down by reason type.** Distribution of (a) total conversation length and (b) remaining messages after the current turn. Boxes show interquartile range, horizontal lines the median, and red diamonds the mean; n denotes the number of annotated turns per category.

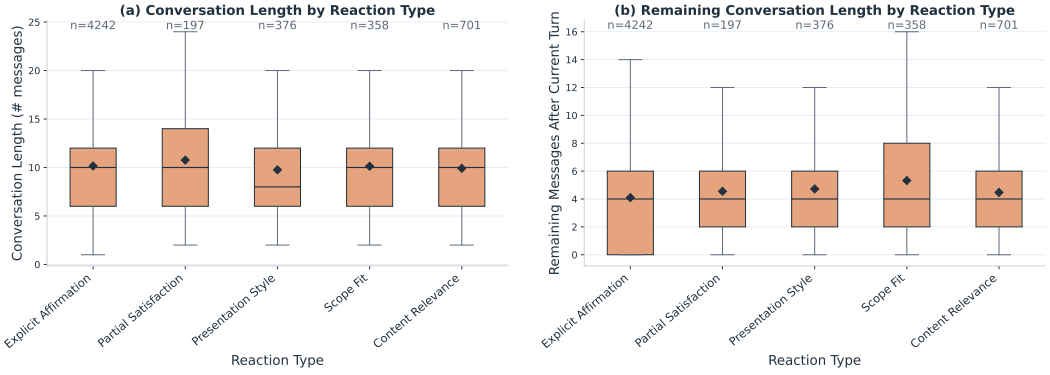


Figure 21: **Conversation length statistics broken down by reaction type.** Distribution of (a) total conversation length and (b) remaining messages after the current turn. Boxes show interquartile range, horizontal lines the median, and red diamonds the mean; n denotes the number of annotated turns per category.

C Details of Data Collection Methodology

C.1 How Thoughts Differ from Intent and Other Theory of Mind Variables

Thoughts in *ThoughtTrace* are free-form, first-person reports of a user's cognitive context at a specific turn. For studying human–AI interaction, our thought taxonomy (Section 4.2) is a better fit than standard Theory of Mind (ToM) variables such as goals, beliefs, and desires (Baker et al., 2009; 2017), owing to two key differences:

(1) Thoughts are turn-based and tied to the interaction. Traditional ToM variables are generally not intended to capture the dynamic reasoning process that unfolds during human–AI interaction. Our taxonomy instead offers a finer-grained characterization of users' turn-by-turn cognitive context: *reasons* capture why a user makes a particular request, and *reactions* capture how the user responds to the assistant at different stages of the interaction.

(2) A single thought spans several ToM variables and reveals information that those variables do not. Consider the reason type *Context Grounding & Constraints*, a common interaction pattern in which the user realizes that the assistant lacks personal or situational context and supplies it. One user reported: *"I live in Japan and forgot to mention this in my initial question, which made the AI give me suggestions that are a little hard or expensive for me to obtain."* This single thought

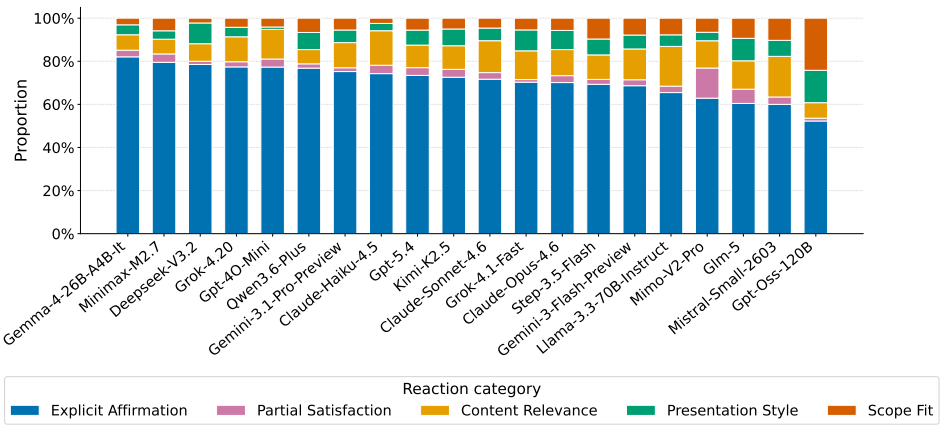


Figure 22: **Distribution of user reaction categories across language models.** Models are sorted by explicit affirmation rate (descending). Each bar represents the proportion of reactions falling into five categories for a given model.

contains a belief about why the previous turn fell short, personal context, and an intent for the next turn; no single ToM variable fully describes it.

Moreover, the categories in our taxonomy describe the context behind subsequent user messages, and the assistance the user needs, more concisely than an enumeration of several mental variables. In the example above, the label *Context Grounding & Constraints* directly indicates why and how the user wants the assistant's response to be revised. These categories also allow us to uncover how human–AI interaction patterns evolve over the course of a conversation, as shown in Thought Property 4 (Section 4.2).

To connect our taxonomy to existing frameworks, Table 8 maps thought types in *ThoughtTrace* to related ToM concepts.

Table 8: **Mapping between thought types in *ThoughtTrace* and related Theory of Mind concepts.**

Thought Types	**Related ToM Concepts**
Reasons: Task Motivation & Goal, Task Reorientation	Goals, intentions, desires
Reasons: Content Expectation, Style Expectation	Goals and desires, grounded in specific content or style preferences
Reasons: Context Grounding & Constraints	Beliefs, knowledge
Reactions: satisfaction (Explicit Affirmation, Partial Satisfaction) and dissatisfaction (Content Relevance, Presentation Style, Scope Fit)	Emotions, appraisal; satisfaction and dissatisfaction are also widely studied signals in NLP, HCI, and LLM alignment (Shi et al., 2024; Jin et al., 2025; Peng et al., 2026; Buening et al., 2026)

C.2 User Consent

We recruit participants through Prolific and compensate them at an hourly rate above the applicable minimum wage. The sample consists of participants who self-report English as one of their fluent languages. Participation is voluntary and self-initiated. Institutional Review Board (IRB) approval was obtained prior to conducting the study.

Participants are redirected to our data collection platform, where they are informed of the study purpose ("investigate how people interact with AI chatbots"), told the study takes approximately 20 minutes, and asked to provide informed consent acknowledging voluntary participation, anonymity, and the right to withdraw. The full consent text is shown below.

User Consent

By selecting the "I Agree - Continue" button below, you acknowledge that:

- You must be at least 18 years old to participate.
- You may decline to answer any or all of the following questions by closing this window in your browser.
- You may decline further participation, at any time, without adverse consequences.
- Your anonymity is assured; the researchers running this study will not receive any of your personal information.

C.3 TUTORIAL

Participants are then guided through a step-by-step tutorial on how to interact naturally with the AI chatbot and record contextually grounded thoughts. The tutorial uses plain language and demos of the chat interface to walk participants through each button and feature. The content of each tutorial page is shown below.

Tutorial

Page 1: Introduction
You will now complete a task using an AI chatbot, a tool you can interact with to assist in your daily activities.

Page 2: Example Task
Plan a trip for yourself. You can plan the trip based on your budget, time, and preference to anywhere in the world—a city, countryside, island, or any destination you choose. Be as creative and realistic as possible. By the end, you should have a complete itinerary.

Guidelines:

- You will have 10 minutes to complete this task.
- When time is up, you'll automatically move on to the next step.

Page 3: Recording Your Thoughts
Please chat as you normally would, but also write down your thoughts during the conversation. Your thoughts help us better understand what users think about the AI chatbot, and we will evaluate their quality. You can write down your thoughts at any time, even after the timeout.

Your thoughts should include:

1. **For each AI response:** Your reactions to the response, including where and why you are satisfied or dissatisfied.
2. **For each of your messages:** Your reasons for sending the message.

Page 4: Chat Interface *[Demo]*
This is exactly what the chat interface will look like. Let's go through each component.

Page 5: Chat Area *[Demo]*
Please chat naturally here. Enter your message in the input box and click the "Send" button.

Page 6: "+ Reasons" Button *[Demo]*
Click this to provide your reasons for sending the message. It will be private and not visible to the AI.

Page 7: "+ Reactions" Button *[Demo]*
Click this to share your reactions to the AI's response. It will be private and not visible to the AI.

Page 8: "New Chat" Button *[Demo]*
You can click the "New Chat" button to start a new chat.

Page 9: "Finish Task" Button *[Demo]*
Once you're done with the task, click the "Finish Task" button.

C.4 Chat Interface

The chat interface is a web application built with HTML, CSS, and JavaScript, backed by Firebase Firestore for real-time data persistence. A screenshot of the interface is shown in Figure 23.

Instruction. A gradient-styled header bar displays the task instruction: *Think of a daily task (e.g., problem-solving, decision-making, planning, creating, brainstorming, or learning) where you would like help from AI. Use the AI chatbot to help complete it.*

Timer. To the right of the header bar, a countdown timer is initialized to 10:00 (600 seconds). The timer pulses with a yellow warning animation when one minute remains. When the timer reaches zero, the text input and send button are disabled, and the placeholder text changes to "Time's up! Please finish the task." Participants may still annotate thoughts after timeout.

Chat Area. The chat area is the main scrollable region where the conversation is displayed. User messages appear right-aligned with a purple gradient background and white text, while assistant messages appear left-aligned with a white background and dark text. Assistant responses are rendered using the marked.js Markdown parser, supporting formatted output. Each message includes a timestamp.

Thought Annotation System. Below each message is a "thought section" containing:

- For **user messages**: a green **"+ Reasons"** button. Clicking it reveals a textarea with the placeholder "Your reasons for sending this message..." along with Save and Cancel buttons. Saved annotations appear as yellow-highlighted cards labeled "your reason" in orange uppercase text.
- For **assistant messages**: a yellow **"+ Reactions"** button. Clicking it reveals a textarea with the placeholder "Your reactions to this response, where and why you are satisfied or dissatisfied..." along with Save and Cancel buttons. Saved annotations appear as yellow-highlighted cards labeled "your reaction".

Multiple thoughts can be attached to a single message. Thoughts are private and not sent to the AI.

Input Area. At the bottom of the chat page, a row contains: (1) a resizable textarea for composing messages, supporting Enter-to-send (Shift+Enter for newlines); (2) a "Send" button; (3) a "New Chat" button that starts a fresh conversation thread while preserving previous threads in the data store; and (4) a "Finish Task" button to submit the current task.

C.5 Optional Thought Annotation

Thought annotation was not required at every turn. Participants were asked to chat naturally with the AI while privately annotating their thoughts, and could add a reason and/or reaction whenever they wished (Appendices C.3 and C.4). Consequently, an empty list of reasons or reactions in *ThoughtTrace* (Appendix C.10) indicates a turn for which the user did not indicate a significant thought, rather than data lost during collection or processing.

This design is intentional for two reasons. First, it avoids forcing participants to fabricate a reason or reaction when they do not have a clear one. Second, it minimizes interruption to the natural interaction and reduces the reactivity caused by thought elicitation (Appendix C.12). Overall, 41.2% of turns are left unannotated, while the remaining 58.8% contain at least one reason or reaction. For example, routine turns such as brief acknowledgments (e.g., "thanks, that works") or simple follow-up clarifications often carry no latent reason or reaction beyond the message itself, and are therefore left unannotated.

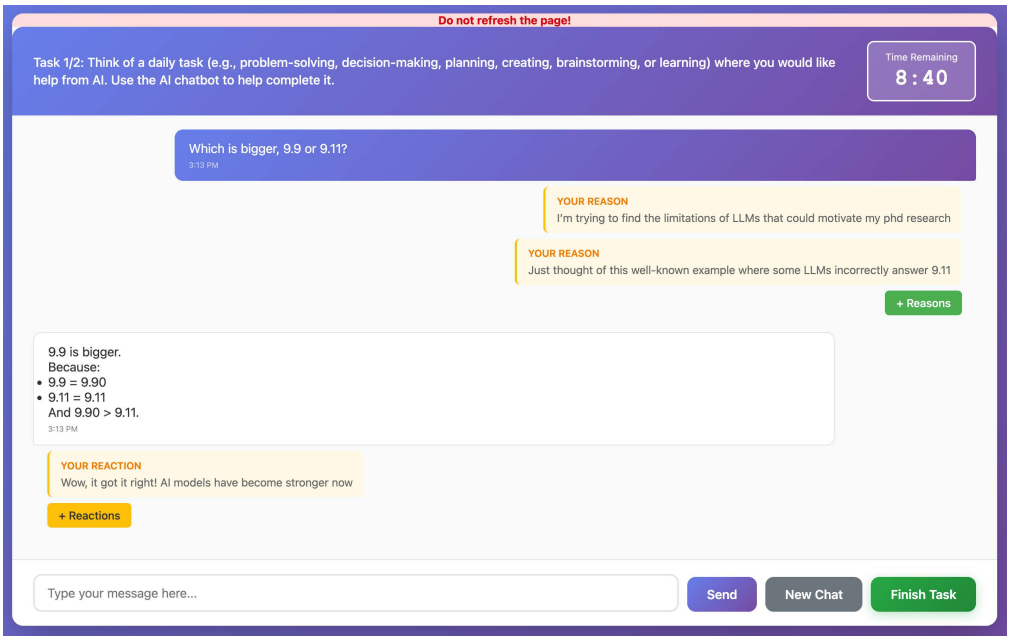


Figure 23: **Chat interface for collecting thought-annotated conversations.** The example shows a participant attaching two reasons to their prompt and a reaction to the assistant's response.

C.6 Post-Chat Surveys

Task Survey. After each task, participants answer the following two open-ended questions:

- What task did you just complete using the AI chatbot?
- In that task, what do you expect from the AI chatbot?

Background Survey. After both tasks, participants complete a demographic survey consisting of the following six questions:

- Age
- Gender (Male / Female / Non-binary / Prefer not to say)
- Education level (High school / Undergraduate / Graduate / Other)
- Occupation (free text)
- Frequency of AI chat usage, measured on a 5-point scale:
  - 1: Never
  - 2: Used a couple of times, but not regularly
  - 3: Once a week
  - 4: Once a day
  - 5: Many times a day
- Main purposes for using AI (free text)

The results of the background survey are summarized in Figure 2 and Section 4.1.

C.7 Quality Control

Because the value of *ThoughtTrace* rests on the quality of its thought annotations, we apply several quality-control mechanisms before, during, and after data collection:

- **Pilot studies.** Before scaling up data collection, we conducted multiple rounds of small-sample pilot studies over several months, which allowed us to iteratively refine the tutorial and platform design.
- **Tutorial and comprehension quiz.** Before data collection, participants complete the guided tutorial (Appendix C.3) and must pass a comprehension quiz demonstrating that they under-

stand how to interact with the chatbot and annotate their thoughts. Participants who fail the quiz are not invited to proceed.
- **Speed-based exclusion.** Prolific flags submissions completed unusually quickly, and we remove such submissions to prevent low-effort participation. The average total completion time, including the tutorial and both tasks, is 30.3 minutes, indicating that participants generally engaged seriously with the study.
- **Post-collection cleaning.** We remove clearly non-serious or malformed sessions while intentionally preserving most naturally occurring variation in the data (Appendix C.8).

C.8 Data Cleaning

After data collection, we retain most of our collected data to preserve its original characteristics, and remove only a small portion, approximately 3% of the raw conversations, in the three cases below:

- In very few cases, Prolific automatically flags participants who complete the task unusually quickly, indicating a lack of serious engagement.
- In very few cases, the chatbot does not respond or responds very slowly, while our system allows users to send multiple messages in the meantime. We remove part or all of such conversations when they contain consecutive user messages and result in strange, low-quality messages or thought annotations.
- In very few cases, we remove extremely low-quality conversations with no thought annotations and incomplete survey responses.

C.9 Collection Cost and Scalability

Our data collection pipeline is designed to make human thought annotation scalable and reproducible: participants are recruited through Prolific, directed to our custom chat platform, trained through a standardized tutorial and quiz, and then privately annotate their own thoughts in situ while completing self-defined tasks (Section 3). Building *ThoughtTrace* at its current scale took approximately six months and cost about $6,000, including multiple rounds of pilot studies, iterative refinement of the collection protocol and interface, participant recruitment, the main data collection, and subsequent quality checks and metadata processing.

Automatically constructing such data at scale remains challenging, which is part of the value of *ThoughtTrace*. Even frontier LLMs struggle to recover users' actual thoughts from conversation context (Thought Property 2 in Section 4.2), and multiple plausible thoughts may be consistent with the same observable interaction. Synthetic thoughts therefore cannot yet substitute for first-person human reports. Instead, *ThoughtTrace* provides previously unavailable ground-truth supervision that can support more scalable approaches, such as training models or user simulators to predict and generate human-like latent states. Existing user-simulation work has largely relied on synthetic internal states because real thought data was unavailable (Wu et al., 2026), and *ThoughtTrace* helps fill this gap.

C.10 Data Format

Each record in *ThoughtTrace* corresponds to a single conversation in which a participant interacted with one of 20 language models to complete an open-ended everyday task. A participant may contribute multiple conversations across two tasks. For each conversation, we record a conversation ID, the model name and provider, the start and last-activity timestamps, a post-hoc task summary and task expectation, and the participant's survey responses (age, gender, education, occupation, AI-usage frequency, and primary use cases).

Each conversation is stored as an ordered list of messages. Each message includes a message ID, timestamp, type (either `user` or `assistant`), message content, and a list of participant thoughts annotated for that message, which is empty if the participant did not annotate that message (Appendix C.5). A thought is either a *reason* attached to a user message or a *reaction* attached to an assistant message. Each thought has its own timestamp, text content, and label, drawn from one of seven reason types or one of five reaction types.

C.11 SAFEGUARDS

ThoughtTrace is released with several safeguards that mitigate the heightened misuse risk of cognitive self-report data. All conversations and annotations were collected under IRB-approved protocols with explicit informed consent, and participants were recruited through Prolific under guarantees of anonymity. No direct identifiers such as names, emails, or contact information appear in the dataset, and only coarse demographic attributes (age, gender, occupation, education, and AI usage frequency) are retained for analysis. We distribute the dataset under a CC-BY-4.0 license intended for research use, and the accompanying dataset card explicitly designates as out-of-scope any attempt to re-identify participants, to build systems that exploit inferred mental states for manipulation or surveillance, or to treat the annotations as a complete record of underlying cognition rather than conscious in-the-moment self-reports. The card also documents known demographic biases and the reactivity inherent to thought elicitation, so that downstream users can apply *ThoughtTrace* within its validated scope of studying latent user thoughts in multi-turn human–AI interaction.

C.12 LIMITATIONS

While *ThoughtTrace* offers a unique window into the thoughts that accompany human–AI interactions, the very act of eliciting such thoughts imposes methodological constraints. We surface three limitations here and explain why each is inherent to in-situ thought collection rather than an artifact of our particular design:

- **Reactivity of thought externalization.** A well-established finding in cognitive science is that asking participants to report on anything beyond their primary task—even after the task is complete—can reshape the task itself. A participant who knows they will later annotate their thoughts may unconsciously adjust their interaction to make those annotations easier to produce: for example, polarizing their stated preferences or adopting cleaner intentions, since extreme or well-defined mental states are easier to articulate than ambiguous ones. This reactivity is fundamentally unavoidable whenever mental states are made explicit: any protocol that renders thoughts observable must also make the participant aware that they are being observed. We therefore interpret the collected annotations as *thoughts-as-reported* rather than *thoughts-as-occurred*, and we design the interface to minimize interruption and framing cues so that reactivity is reduced, though it cannot be eliminated.
- **Conscious versus subconscious cognition.** Externalized thoughts capture only those mental states that participants can consciously access and verbalize. Decades of work in psychology and behavioral science show that a substantial share of human behavior is shaped by subconscious processes, implicit associations, and automatic judgments that elude verbal report. As a consequence, *ThoughtTrace* should be read as a record of users' *explicit* reasoning about their interactions, not as a complete account of the cognitive processes driving them. We make this scope explicit here, and we encourage downstream users of the dataset to treat annotations as a conscious overlay on, rather than a transcript of, the underlying cognition. We view this as a scoping decision rather than a deficiency: consciously articulated thoughts are themselves a signal that existing interaction datasets do not provide.
- **Recruited rather than fully in-the-wild participants.** Although our goal is to characterize thoughts during naturalistic human–AI interactions, participants are recruited through Prolific rather than drawn from unsolicited model traffic. This is a practical necessity: users of a public model/API service have no incentive to annotate their thoughts, and truly unsolicited thought collection would require invasive instrumentation that is neither ethical nor feasible at scale. Recruitment therefore introduces a modest selection effect. Reassuringly, however, our demographic analysis (Section 4.1) shows that *ThoughtTrace* reflects a diverse spectrum of AI users and everyday use cases, consistent with the profile of frequent AI users in the real world, suggesting that the recruitment-induced distribution shift is small relative to the value of obtaining rich, consented thought annotations at scale.

D Details of Analyses and Experiments

D.1 Conversation Property 1: *ThoughtTrace* Captures a Representative Spectrum of Users

To characterize the participant pool behind *ThoughtTrace*, we extract self-reported demographic and usage information from the post-task survey completed by each annotator alongside their conversations. For each participant, we retain the first survey response and aggregate responses along six axes: *age*, *gender*, *education*, *occupation*, self-reported *frequency* of LLM use, and free-text *purposes* of use.

Age is parsed as an integer and grouped into canonical brackets: 18–24, 25–34, 35–44, 45–54, 55–64, and 65+. Usage frequency is mapped from a 1–5 Likert scale to human-readable anchors ranging from "Never" to "Many times a day." Gender and education are mapped to fixed category sets, including Male/Female/Non-binary for gender and High school/Undergraduate/Graduate/Other for education. For the two open-ended fields, *occupation* and *purposes*, we canonicalize responses by stripping whitespace and punctuation, lowercasing for deduplication, and re-casing labels for display. Purposes are further grouped into a small set of semantically coherent categories, including *Learning*, *Working*, *Brainstorming*, *Research*, etc., using keyword-based rules. Any unmatched responses are retained under their title-cased surface forms.

We compute counts for each group. Fixed-category axes are sorted by descending frequency with a deterministic tiebreaker, while open-ended axes are limited to the top eight entries whose display labels fit within a fixed character budget. The resulting statistics are rendered as a single six-panel horizontal bar chart, with one panel per demographic axis.

D.2 Conversation Property 2: *ThoughtTrace* Features Long-Horizon, Topically Diverse Conversations

Conversation and Message Lengths. These analyses build on a shared message-level data frame, computed over every conversation in *ThoughtTrace*, that tags each message with its role (user or assistant), records its one-indexed turn position, and counts tokens with the tiktoken encoding for GPT-4o. The token-level conversation length analysis (*ThoughtTrace* vs. WildChat and LMSYS-Chat-1M) aggregates this frame into per-conversation token totals for *ThoughtTrace*, and obtains matching totals for WildChat (and likewise for LMSYS-Chat-1M) by counting tokens across every message of WildChat-1M using the same tiktoken encoder, with a whitespace-based regex as a fallback. All three corpora are then bucketed into fixed 1,000-token bins (0–1k, 1k–2k, ..., 14k–15k; conversations above the cap fold into the last bin), converted to percentages of conversations per bin, and rendered as side-by-side bars.

The turn-level conversation length analysis (*ThoughtTrace* vs. WildChat and LMSYS-Chat-1M) follows a parallel structure at the turn level. It derives the total turn count of each *ThoughtTrace* conversation by taking the maximum turn position per conversation, and it obtains WildChat (and likewise LMSYS-Chat-1M) turn counts by enumerating the conversation field of every WildChat-1M row. All three corpora are bucketed into even bins, normalized to percentages of their respective corpora, and drawn as side-by-side bars centered on the even integers. The x-axis is limited to [1, 30].

The *Prompt and Response Lengths* analysis consumes the shared data frame directly and partitions token counts by role. It bins counts at a width of 200 tokens up to a cap of 4,000 and overlays two histograms on a single axis, with user prompts in blue and assistant responses in pink. Dashed vertical lines mark the per-role median token count; the y-axis uses a thousands formatter (e.g., "1k").

The *Prompt and Response Length by Turn Position* analysis also reuses the shared frame, restricts it to turn positions 1 through 20, and takes the mean token count within each (role, turn position) cell. To respect the alternating structure of the dialogue, user means are retained only at odd positions, and assistant means are retained only at even positions. The two sequences appear as line plots with circular markers on a shared axis; the x-axis ticks span 1–20, and the y-axis reports average tokens per message.

Conversation Topics. We label conversation topics using an LLM (GPT-5.4) with a predefined topic taxonomy to assign all topics clearly present in each conversation, rather than forcing a single primary label. For each conversation, we concatenate the user and assistant turns into a single transcript and prompt the model with the full taxonomy and labeling instructions, requesting a JSON response containing the relevant taxonomy labels. The model is called with temperature 0 for deterministic outputs, and the returned labels are deduplicated and validated against the allowed taxonomy list via a cleaning step that discards any hallucinated or out-of-taxonomy labels.

This multi-label design allows a single conversation to be tagged with multiple topics when it spans several domains—for instance, a conversation touching on both programming and education receives both labels. After labeling, we aggregate topic counts across all conversations and organize them into a two-level hierarchy: topics are grouped into broader categories (e.g., "Technology," "Business & Society," "Arts & Entertainment") defined by a manual grouping, with any topics not covered by these predefined groups collected under "Other Topics." This hierarchical structure is then visualized as a donut chart of parent categories (Figure 3(b)) and a per-subtopic bar chart (Figure 13), providing an at-a-glance view of the topical distribution across the dataset.

We provide the topic labeling instructions for the LLM below.

Conversation Topic Labeling Instruction

Label **all** topics that are clearly present in the content of the current conversation turn, or that are essential to completing the user's task. Many turns involve **multiple overlapping domains**, and **you should apply every relevant topic label**, not just the most obvious one. For example, a creative writing task about nature and exploring new lands should be labeled with "Literature & Writing", "Nature & Environment", and "Travel & Tourism". Also, a query asking to build a website for a fashion business should be labeled with "Fashion & Beauty", "Business & Finances" & "Technology, Software & Computing". Include **all relevant topics**, especially when the content spans more than one thematic area. Avoid under-labeling.

We provide the corresponding topic taxonomy below.

Conversation Topic Taxonomy

- **Adult & Illicit Content:** Content involving mature or age-restricted themes, including topics related to sex, drugs, or alcohol.
- **Art & Design:** Topics about visual art, creative techniques, aesthetics, or design principles.
- **Business & Finances:** Topics involving any type of companies, markets, business practices, management, or personal financial planning.
- **Culture:** Topics concerning traditions, customs, social norms, lifestyle identities, or cultural critique.
- **Economics:** Topics involving macro- or microeconomic systems, financial theory, trade, inflation, or economic policy.
- **Education:** Topics related to teaching, learning, academic subjects, school systems, or knowledge transmission. Do not multi-label with 'Technology, Software & Computing' unless the task is about technical content (e.g., teaching code or explaining software).
- **Employment & Hiring:** Topics about careers, job applications, resumes, hiring processes, or workplace dynamics.
- **Entertainment, Hobbies & Leisure:** Topics related to movies, television, music, games, crafts, or recreational activities.
- **Fantasy / Fiction / Fanfiction:** Creative or fictional writing that includes imaginary worlds, character scenarios, or fan-authored extensions of media.
- **Fashion & Beauty:** Topics about personal style, clothing trends, beauty products, grooming, or industry standards.
- **Food & Dining:** Topics involving cuisines, recipes, cooking, restaurants, or dietary habits.
- **Geography:** Topics about physical locations, world regions, maps, or geopolitical features.

- **Health & Medicine:** Topics involving physical or mental health, medical knowledge, treatments, or wellbeing.
- **History:** Topics related to past events, timelines, historical figures, or historical analysis.
- **Housing:** Topics about real estate, renting, home ownership, architecture, or urban planning.
- **Immigration / Migration:** Topics about cross-border movement, cultural adaptation, visas, or diaspora experiences.
- **Insurance & Social Scoring:** Topics related to insurance policies, risk models, or institutional scoring systems (e.g., credit/social scores).
- **Interpersonal Relationships & Communication:** Topics involving any form of signal where people interact with one another in personal, social, or professional settings. This includes: (1) Romantic, familial, platonic relationships, friendship, dating, breakups, intimacy, or emotional connection, (2) Communication advice (e.g., how to respond, how to ask, how to apologize), (3) Conflict resolution, emotional support, or interpersonal misunderstandings, (4) Social etiquette, small talk, expressing emotions or affection, tone-setting, or conversational timing. Label this even when the focus is not only when it's on the relationship itself, but on how to say something or respond in a conversation.
- **Law, Criminal Justice, Law Enforcement:** Topics about legal systems, crime, law enforcement, policing, or justice procedures.
- **Lifestyle:** Topics about routines, productivity, wellness, habits, or personal philosophies of daily living.
- **Linguistics & Languages:** Topics involving grammar, syntax, language families, translation, or linguistic theory.
- **Literature & Writing:** Topics about books, authorship, literary analysis, storytelling techniques, or writing practices.
- **Math & Sciences:** Topics about quantitative reasoning, scientific disciplines (e.g., math, physics, biology, chemistry), or scientific inquiry.
- **Nature & Environment:** Topics related to natural ecosystems, wildlife, conservation, weather, or environmental issues.
- **News & Current Affairs:** Topics referencing current or recent events, media coverage, or public opinion.
- **Non-software Engineering & Infrastructure:** Topics related to mechanical, civil, or structural engineering, and infrastructure systems (e.g., bridges, waterworks).
- **Politics & Elections:** Topics about political ideologies, government systems, elections, politicians, or civic processes.
- **Psychology, Philosophy & Human Behavior:** Topics involving cognitive science, behavioral analysis, mental health, philosophical reasoning, or ethics.
- **Religion & Spirituality:** Topics involving religious traditions, beliefs, practices, spiritual experiences, or theology.
- **Social Issues & Movements:** Topics involving inequality, advocacy, discrimination, social justice, or civil movements.
- **Sports:** Topics about teams, players, sporting events, fitness, or athletics.
- **Technology, Software & Computing:** Topics involving computers, software, hardware, internet systems, or technological innovation. This includes: Programming languages, code generation, debugging, or algorithms, Software tools (e.g., Excel, Photoshop), platforms, or operating systems, Artificial intelligence, machine learning, or large language models, Consumer tech (e.g., smartphones, laptops, smart devices), Networking, cybersecurity, web development, or IT systems, Tech industry news, product comparisons, or emerging technologies.
- **Transportation:** Topics involving travel modes, infrastructure, public transportation, or logistics.
- **Travel & Tourism:** Topics about travel destinations, itineraries, tips, or exploration.
- **Video Games:** Topics involving game design, gameplay, genres, or gaming culture.
- **Other:** Any topic not fitting the categories above.

D.3 Conversation Property 3: *ThoughtTrace* Conversations Are Dominated by Task Extension

We analyze conversational structure by using an LLM (GPT-5.4) to label the relationship between each user turn and the immediately preceding user turn. For each conversation, we extract all user turns in order; the first turn is automatically labeled as "First request," and for every subsequent turn, we prompt the model with both the previous and current user prompts, asking it to classify their relationship using a predefined taxonomy. The model is called with temperature 0 for deterministic outputs and returns a single JSON label, which is then normalized against the allowed taxonomy via a cleaning function that uses case-insensitive matching and keyword-based fallback rules to handle minor variations in the model's output.

This turn-level labeling assigns each user message one of four relationship types relative to its predecessor: (1) Extend, deepen, or build on the prior task, (2) Re-attempt or revise the prior task, (3) New variation of the prior task, or (4) Completely new request. The resulting sequence of relationship labels is stored both at the conversation level and attached directly to each individual user message, enabling fine-grained analysis of how users navigate within a conversation, whether they primarily continue and elaborate on a task, revise their prior attempt, explore variations, or shift to an entirely different request. This captures the structural dynamics of multi-turn interactions beyond what topic labels alone reveal.

We provide the labeling instructions for the LLM below.

Message Multi-turn Relationship Labeling Instruction

Label one type of relationship that is present between the previous user prompt and the current user prompt in the conversation, using **exactly one option** from the list below.

We provide the corresponding multi-turn relationship taxonomy below.

Message Multi-turn Relationship Taxonomy

- **First request:** The initial input or prompt provided by the user to start a new conversation thread.
- **Completely new request:** A user input that shifts to a different topic or context, unrelated to the previous turns.
- **Re-attempt/revision on prior task:** A follow-up input where the user revisits or retries a previously attempted task, possibly from earlier in the previous conversation. This often includes revised phrasing, clarification, or corrections after perceived issues, dissatisfaction, or misunderstanding.
- **New variation of prior task:** A follow-up input that is a related task to the prior turn, but explores a different angle or formulation of a previously attempted task.
- **Extend, deepen, or build on prior task:** A continuation that elaborates on, adds complexity to, or builds logically from the prior turn, indicating that the previous response was useful but not complete.

D.4 Thought Property 1: Thoughts Are Different from Messages

To quantify how much of a user's underlying thinking is already reflected in their visible utterance, we measure the semantic coverage between user messages and their associated thoughts. For each eligible user turn, we extract the user's reason (their stated motivation for sending the current message) and the user's reaction (their response to the previous assistant message). We then use an LLM (GPT-5.4) to score how well the user's utterance conveys each type of thought on a 1–5 scale, where 1 indicates no meaningful overlap and 5 indicates full coverage. The scoring follows a structured rubric: the model receives the utterance and the thought as input and returns a single integer representing the degree of semantic overlap.

This analysis reveals the extent to which thoughts provide information that differs from what users explicitly express in their messages. A low average coverage score suggests that the thoughts capture latent user intent and reactions that are largely missing from the surface-level utterance, supporting the claim that thoughts constitute a meaningfully distinct signal from the conversation text alone.

By evaluating reason coverage and reaction coverage separately, we can further distinguish whether users tend to omit their motivations for a new request versus their evaluative responses to prior assistant outputs, offering a more nuanced understanding of where the gap between utterances and internal reasoning is most pronounced.

We provide the prompt used to score how well the user's utterance conveys their thought below.

Semantic Coverage Evaluation Prompt

System Prompt:
You are a semantic coverage evaluator. Given a user's utterance and their underlying thought (their reactions to previous assistant responses and/or their reasons for sending the message), score how well the utterance expresses the thought's semantic content.

Scoring rubric:

- **1** – No meaningful overlap; the utterance misses or contradicts the thought
- **2** – Minimal overlap; only a vague or incidental connection
- **3** – Partial coverage; the utterance captures some but not the core of the thought
- **4** – Good coverage; most of the thought is conveyed, with minor gaps
- **5** – Full coverage; the utterance fully and accurately expresses the thought

Output only a single integer (1–5). No explanation.

User Prompt:
Utterance: {user_message}
Thought: {thought}

Score:

D.5 Thought Property 2: Thoughts Are Difficult for LLMs to Infer

To assess how difficult it is to recover user thoughts from surface dialogue context, we prompt three frontier models—GPT-5.4, Gemini 3.1 Pro Preview, and Claude Opus 4.6—to infer two types of thoughts: (1) the user's *reason* for sending their most recent message and (2) the user's *reaction* to the assistant's most recent response. These two subtasks are conditioned on different dialogue contexts. For *reasons*, we provide the conversation history up to and including the target user turn. For *reactions*, we provide the conversation history up to and including the assistant turn being reacted to, optionally followed by the subsequent user message when available, since this follow-up often provides the strongest signal about whether the assistant's response satisfied the user. Both prompts consist of a system message that specifies the predictor's role and constrains the output to a single sentence in the user's voice, followed by a user message containing the formatted dialogue context.

Each model prediction is compared against the corresponding human-written thought using an LLM-as-a-judge. To mitigate self-preference bias, we deliberately use a judge model different from the predictor: predictions from GPT-5.4 are judged by a random choice between the two non-OpenAI models, and predictions from each non-OpenAI model are judged by GPT-5.4. The judge follows a fixed five-point rubric, ranging from 1 for no meaningful overlap or contradiction to 5 for a full semantic match while ignoring surface wording. We parse the judge's response as an integer and clamp it to the range $[1, 5]$. All predictions and judgments are cached to disk, and we report per-model averages as well as the unweighted mean across the three predictors for each thought type.

This analysis tests a key assumption: if thoughts were simply recoverable from the observable conversation, they would add little value as annotations. A low average similarity score between predicted and actual thoughts suggests that even a capable language model, given full conversational context, cannot reliably reconstruct what users are actually thinking, whether that concerns their motivations for a request or their evaluative responses to assistant outputs. By evaluating reasons and reactions separately, we can further identify which type of thought is harder to infer, revealing where the gap between observable dialogue and latent user cognition is most pronounced. Together with the coverage analysis, these results demonstrate that thoughts constitute a genuinely novel signal that is both distinct from user utterances and difficult to recover from context.

We provide the prompt used to infer the users' reasons below.

Reason Inference Prompt

System Prompt:
You are a user intent predictor. Given a conversation context, infer the user's reason for sending the most recent message. Express the reason from the user's point of view. Output only a single sentence. No explanation.

User Prompt:
Conversation context: {context}

Predicted reason:

We provide the prompt used to infer the users' reactions below.

Reaction Inference Prompt

System Prompt:
You are a user reaction predictor. Given a conversation context, infer the user's reaction to the most recent assistant message—identifying where and why they are satisfied or dissatisfied. Express the reaction from the user's point of view. Output only a single sentence. No explanation.

User Prompt:
Conversation context: {context}

Predicted reaction:

We provide the prompt used to evaluate the predicted reasons against the actual human-annotated reasons below.

Reason Semantic Similarity Evaluation Prompt

System Prompt:
You are a semantic similarity evaluator. Given two reasons describing why a user sent a message in a conversation, score how well the predicted reason matches the actual reason in semantic meaning (ignore surface wording). Scoring rubric:

- **1** – No meaningful overlap; the reasons disagree or contradict each other
- **2** – Minimal overlap; only a vague or incidental connection
- **3** – Partial match; the predicted reason captures some but not the core of the actual reason
- **4** – Good match; most of the actual reason is reflected, with minor gaps
- **5** – Full match; the predicted reason fully and accurately conveys the same meaning as the actual reason

Output only a single integer (1–5). No explanation.

User Prompt:
Actual reason: {actual}
Predicted reason: {predicted}

Score:

We provide the prompt used to evaluate the predicted reactions against the actual human-annotated reactions below.

Reaction Semantic Similarity Evaluation Prompt

System Prompt:
You are a semantic similarity evaluator. Given two reactions describing a user's response to an assistant's message in a conversation, score how well the predicted reaction matches the actual reaction in semantic meaning (ignore surface wording). Scoring rubric:

- **1** – No meaningful overlap; the reactions disagree or contradict each other
- **2** – Minimal overlap; only a vague or incidental connection
- **3** – Partial match; the predicted reaction captures some but not the core of the actual reaction
- **4** – Good match; most of the actual reaction is reflected, with minor gaps
- **5** – Full match; the predicted reaction fully and accurately conveys the same meaning as the actual reaction

Output only a single integer (1–5). No explanation.

User Prompt:
Actual reaction: {actual}
Predicted reaction: {predicted}

Score:

D.6 Thought Property 3: Thoughts Are Diverse in Content

To categorize reasons and reactions into fine-grained types, we use `GPT-5.4` to assign labels from a predefined taxonomy. The prompting setup is tailored to the distinct contextual nature of each thought type. For labeling *reasons*, we provide the conversation history up to and including the current user message, followed by the target reason text to label. This context enables the model to interpret the underlying motivation for a user utterance in light of prior turns. We preserve the dialogue structure but do not include the full content of assistant responses. This design reflects that the user's intent is primarily expressed through their own sequence of actions across turns, rather than the specific wording of assistant replies. By focusing on user-side signals while retaining conversational structure, the model is better guided to infer why a particular message was produced.

In contrast, *reactions* are inherently localized: the model receives only the single assistant response that the user reacts to, followed by the corresponding reaction text. This is because a reaction reflects the user's immediate evaluation of a specific response, and is therefore primarily determined by the content and presentation of that response itself.

Reasons Labeling Prompt

System Prompt:
You are a thought classification expert. Your job is to read a conversation between a human and an AI assistant, then classify the internal thought the human had during that conversation into exactly one of the categories below.

Categories:

- **task_motivation**: The user introduces the task and explains their underlying goal, need, or real-world motivation for initiating the conversation.
  Example: "I chose this because I am preparing for USMLE and need a study plan."
- **task_continuation**: The user engages in follow-up interaction to refine, expand, or probe deeper into the current task without changing the overall goal.
  Example: "trying to conclude" / "to develop previous message."
- **task_reorientation**: The user changes their objective or redirects the conversation toward a different task or outcome.
  Example: "Changing plan." / "changing my search from common questions to totally differents one and check it is work on feed or do real sreach"
- **content_expectation**: The user specifies what the response should achieve in terms of substance, such as level of detail, correctness, or type of information.
  Example: "i want him to do a pretty general model so i can change a bit and be good to go"
- **style_expectation**: The user defines how the response should be structured or presented, including format, tone, or organization.
  Example: "I asked this to get a more detailed and practical plan that I can actually follow." / "I won't read so much"

- **context_grounding_and_constraints**: The user provides situational context, personal preferences, or constraints that shape the solution space.
  Example: "I wanted to tell the AI all that I needed to get done that is relevant to my day by including all of the important things I need to get done today."
- **social_and_others**: The user engages in interaction management or non-task-oriented communication such as greetings, acknowledgments, or meta-comments.
  Example: "Hi!" / "Thanks, that helps." / "AI convince me"

Instructions:
- Analyze the given thought text carefully
- Classify it into exactly one category
- Choose the category that best represents the primary intent of the thought
- If a thought fits multiple categories, select the most dominant/primary one
- Be precise and consistent in your classifications

Return only the category label in lowercase with underscores. No explanation.

User Prompt:
Conversation:

```
Human message: ...
AI assistant response: ...
...
Human message: ...
Human thoughts for this message: {reason}
```

Label:

Reactions Labeling Prompt

System Prompt:
You are a reaction classification expert. Your job is to read an AI assistant response and the internal reaction the human had after reading it, then classify that reaction into exactly one of the categories below.

Categories:
- **explicit_affirmation**: The user produces clear positive signals in text, expressing that the response fully met their expectations.
  Example: "Perfect.", "That's exactly it."
- **partial_satisfaction**: The user acknowledges the output positively but immediately requests a minor tweak.
  Example: "This is great, just make it a bit shorter."
- **presentation_style**: The user expresses dissatisfaction with how the response is delivered, including tone, wording, structure, or overall presentation.
  Example: "The email is okay, but it sounds too serious and uses big words."
- **scope_fit**: The user expresses dissatisfaction with how much or how broadly the response is provided, including its length, level of detail, and breadth.
  Example: "Way too detailed.", "Too much words.", "Too long."
- **content_relevance**: The user expresses dissatisfaction with what is included in the response, including missing information or irrelevant material.
  Example: "The response is clearer now, but I would like suggestions for specific hotels within my budget."

Instructions:
- Analyze the given reaction text carefully
- Classify it into exactly one category
- Select the most dominant intent if multiple categories apply
- Be precise and consistent

Output only the category label in lowercase with underscores. No explanation.

User Prompt:
`AI assistant response: {response}`
`Human reaction to this AI response: {reaction}`
Label:

D.7 Thought Property 4: Thought Dynamics Depend on Conversation Stages

Relationship to conversation stage. To characterize how user thoughts evolve over the course of a conversation, we construct Sankey-style flow visualizations over four normalized dialogue stages, *Early*, *Mid-Early*, *Mid-Late*, and *Late*, corresponding to normalized positions of 0%, 33%, 67%, and 100% of the conversation. For each stage, annotated labels are aggregated and normalized to obtain category-level percentage distributions, where stacked vertical bars represent the relative frequency of each category at a given stage. To model transitions across stages, we estimate pairwise flows between categories in consecutive stages. For a source category c_i at stage t, its outgoing mass is distributed across categories at stage $t + 1$ proportionally to the target-stage category frequencies, yielding a dense transition matrix while preserving the total mass associated with each source category. The resulting flows are rendered as smooth ribbons connecting stacked segments across stages, enabling a compact visualization of temporal shifts in conversational patterns.

Relationship to conversation's topic and message's multi-turn relationships. To examine how thought types vary across conversational contexts, we constructed cross-tabulation heatmaps between thought labels and two categorical dimensions: conversation topic and multi-turn relationship type. For reason labels, we paired each labeled reason with both the conversation-level topic annotations and the message-level multi-turn relationship label assigned to that same user turn. For reaction labels, we paired each labeled reaction with the conversation-level topic and, crucially, with the multi-turn relationship label of the *next* user message rather than the current one, capturing the forward-looking relationship between an assistant's response characteristics and the user's subsequent behavioral choice. All heatmaps were computed as normalized percentages. To support analysis at multiple granularities, we optionally aggregate the 36 individual topic labels into 7 broader thematic groups (e.g., Technology, Business & Society, Health & Relationships), using a predefined mapping consistent with the topic hierarchy defined in *Conversation Property 2*.

Relationship to conversation length. To assess how thought types relate to conversation structure, we computed two positional statistics for each thought label: total conversation length (number of messages in the conversation where the thought occurs) and remaining conversation length (number of messages after the current turn). These were collected separately for reason labels on user messages and reaction labels on assistant messages. The resulting distributions were visualized as paired box plots with overlaid mean markers, enabling comparison of both central tendency and spread across thought types. Sample sizes were annotated on each box to contextualize the statistical reliability of each category. This design reveals whether certain thought types tend to appear in longer or shorter conversations and whether they cluster toward the beginning or end of a conversational session.

D.8 Thought Utility 1: Thoughts Predict User Behavior

This section details the next-message prediction experiment in Section 5.1: dataset filtering, prediction prompts for the history-only and thought-augmented conditions, and the semantic similarity scoring protocol.

To evaluate whether thought annotations improve the ability to anticipate user behavior, we conduct a next-message prediction experiment. For each assistant message followed by a user turn, we construct two versions of the conversation context: a *history-only* version containing only the raw dialogue history, and a *thought-augmented* version that interleaves the user's annotated reasons and reactions at the appropriate turns. We restrict the evaluation to examples whose thought annotations are high-quality, i.e., substantive and informative about the user's latent intent or attitude beyond what is already evident in the conversation surface. Concretely, we use an LLM judge to rate every thought annotation on a 1–5 quality scale and keep only examples scored ≥ 4, ensuring that the comparison reflects the value of genuinely informative thoughts rather than boilerplate filler. We

then prompt six LLM predictors, GPT-5.4, Gemini 3.1 Pro Preview, Claude Opus 4.6, Kimi-K3, GLM-5.2, and Muse Spark 1.1, to predict the user's next message under each condition independently. We assign each prediction a semantic similarity score in $[0, 100]$ relative to the actual next user message, using an LLM judge with the prompt shown below. To avoid self-evaluation bias, each predictor's outputs are scored by a judge sampled uniformly at random from the other five models.

This setup directly quantifies the predictive utility of thought annotations: a higher thought-augmented similarity relative to the history-only baseline indicates that knowing what users are thinking provides an actionable signal for anticipating their subsequent messages beyond what the conversation surface alone reveals. Across all six predictor models (Table 1), thought-augmented prediction consistently outperforms history-only prediction, suggesting that thoughts capture information that is novel, hard to recover, and practically valuable for modeling user behavior in multi-turn conversations.

We provide the prompt used to predict the next message in the history-only setting below.

User Message Prediction Prompt (History-only)

System Prompt:
You are a next-message predictor. Given a conversation context, predict the user's next message.

Output only a single likely next user message. No explanation.

User Prompt:
Conversation context: {context}

Predicted next message:

We provide the prompt used to predict the next message in the thought-augmented setting below.

User Message Prediction Prompt (Thought-augmented)

System Prompt:
You are a next-message predictor. Given a conversation context that includes thought annotations, predict the user's next message.

Output only a single likely next user message. No explanation.

User Prompt:
Conversation context with thoughts: {context_with_thoughts}

Predicted next message:

The thought-augmented context uses the same multi-turn dialogue history as the history-only context, but interleaves the user's free-form reason and reaction annotations immediately after the corresponding messages. Multiple reasons or reactions for the same message are joined with "|" and added as labeled "Note:" lines. The resulting transcript is provided directly to the predictor model. The history-only context is identical except that these thought annotations are omitted. We show the format below.

Thought-Augmented Conversation Context Format

User: {user_turn_1}
Note: The user's reasons for sending this message are: {reason_A} | {reason_B}
Assistant: {assistant_turn_1}
Note: The user's reactions to the assistant's message are: {reaction_A} | {reaction_B}
User: {user_turn_2}
Note: The user's reasons for sending this message are: {reason_C}
...

We provide the prompt used to score the semantic similarity between a predicted next message and the actual next message below.

Next-Message Semantic Similarity Prompt

System Prompt:
You are a semantic similarity evaluator. Given an actual user message and a predicted next user message, score how well the prediction matches the actual message in semantic meaning (ignore surface wording, length, and style).

Scoring rubric (continuous scale):
0 - No meaningful overlap; the prediction disagrees with or is unrelated to the actual message
25 - Minimal overlap; only a vague or incidental connection
50 - Partial match; captures some but not the core of the actual message
75 - Good match; most of the actual message is reflected, with minor gaps
100 - Full match; fully and accurately conveys the same meaning as the actual message

Output only a single number between 0 and 100. No explanation.

User Prompt:
Actual next message: {actual}
Predicted next message: {predicted}

Score:

D.9 Thought Utility 2: Thoughts Improve Model Alignment

This section details the alignment experiments in Section 5.2: training data construction for thought-guided and message-guided rewrites, the rewrite prompts, and training and evaluation setups.

Training data. We generate the training data for thought-guided rewrites as follows:

1. **Load and filter conversations:** We load the dataset and retain only conversations with 2–20 turns.
2. **Collect dissatisfaction reactions:** We scan all user reactions labeled as "content relevance", "presentation style", or "scope fit", the three dissatisfaction types defined in *ThoughtTrace*. Each reaction's text serves as the "thought" that guides the rewrite.
3. **Filter to meaningful thoughts:** We discard thoughts that are empty, shorter than six words, or contain no alphabetic characters, ensuring the rewriter has sufficient signal to act on.
4. **Build multi-turn context:** For each remaining candidate, we slice the conversation up to (but not including) the dissatisfying assistant response, yielding a {role, content} message list that ends with the triggering user prompt.
5. **Generate thought-guided rewrites:** We prompt GPT-5.4 with the context, the original response, the dissatisfaction label and its description, and the user's thought, requesting a revised assistant response that addresses the complaint.
6. **Save as DPO pairs:** We store the training data in the standard DPO schema: *prompt* (the multi-turn context up through the triggering user message), *chosen* (the thought-guided rewrite), and *rejected* (the unsatisfactory assistant response from the original dataset).

We generate the training data for message-guided rewrites as follows:

1. **Load and filter conversations:** We load the dataset and retain only conversations with 2–20 turns.
2. **LLM-classify dissatisfaction:** We prompt GPT-5.4 with each (assistant response, user follow-up) pair and ask it to output exactly dissatisfied or satisfied. Each follow-up message classified as dissatisfied serves as the signal that guides the rewrite.
3. **Filter to meaningful messages:** We discard messages that are empty, shorter than six words, or contain no alphabetic characters, ensuring the rewriter has sufficient signal to act on.
4. **Build multi-turn context:** For each remaining candidate, we slice the conversation up to (but not including) the dissatisfying assistant response, yielding a {role, content} message list that ends with the triggering user prompt.

5. **Generate message-guided rewrites**: We prompt GPT-5.4 with the context, original response, and the user's follow-up message, asking for a revised response that preemptively addresses the follow-up so the user would not have needed to push back.
6. **Save as DPO pairs:** We store the training data in the standard DPO schema: *prompt* (the multi-turn context up through the triggering user message), *chosen* (the message-guided rewrite), and *rejected* (the unsatisfactory assistant response from the original dataset).

In both cases, each unsatisfactory assistant response in a multi-turn conversation yields one DPO pair: the preceding conversation serves as the prompt, the original response as the rejected response, and its rewrite as the chosen response. For example, if the user is dissatisfied with the second assistant turn, the resulting pair is constructed as follows.

Example DPO Pair Construction

Prompt: (User Turn 1, Assistant Turn 1, User Turn 2)
Rejected: Assistant Turn 2
Chosen: Thought-guided (or message-guided) rewrite of Assistant Turn 2

The training data sizes for the three training runs are:

1. 1,000 instances using thought-guided rewrites on *ThoughtTrace*, derived from 1,985 conversations (92% of all *ThoughtTrace* conversations).
2. 450 instances using message-guided rewrites on *ThoughtTrace*, derived from the same 1,985 conversations as (1). The smaller size is intentional: it ensures a fair comparison on identical conversations and supports our claim that thoughts surface more dissatisfaction instances than messages.
3. 1,000 instances using message-guided rewrites on WildChat, derived from 4,669 conversations. We process WildChat conversations in random order until we obtain 1,000 filtered instances, matching the size in (1).

Prompts Used. We provide the prompts used to generate the thought-guided and message-guided rewrites below.

Thought-Guided Rewrite Prompt

System Prompt:
You are an expert assistant that rewrites prior assistant responses so they better satisfy the user. You will be given the multi-turn conversation so far, the original assistant response the user was dissatisfied with, and the user's internal "thought" explaining why they were dissatisfied. Your job is to produce a single revised assistant response that directly addresses the user's feedback. Return ONLY the revised response text, with no preambles, no explanations, no meta-commentary, and no markdown fences.

User Prompt:
Conversation context: {context}

Original assistant response (the user was dissatisfied with this): {assistant_response}

User's dissatisfaction label: {label} ({label_description})

User's internal thought about why they were dissatisfied: {thought}

Message-Guided Rewrite Prompt

System Prompt:
You are an expert assistant that rewrites prior assistant responses so they better satisfy the user. You will be given the multi-turn conversation so far, the original assistant response the user was dissatisfied with, and the user's next message which expresses (explicitly or implicitly) what they actually wanted. Your job is to produce a single revised assistant response, issued at the same point as the original assistant response, that proactively addresses the concerns surfaced by the user's next message, so that the user would not have needed to

push back. Return ONLY the revised response text, with no preambles, no explanations, no meta-commentary, and no markdown fences.

User Prompt:
Conversation context: {context}

Original assistant response (the user was dissatisfied with this): {assistant_response}

User's next message (revealing what they actually wanted / why they were unhappy): {user_followup}

Training details. We initialize all models from Qwen3.5-4B (Yang et al., 2025). We conduct Direct Preference Optimization (DPO) training using the Tinker APIs. Across all three experiments, we use a batch size of 64, a learning rate of 1×10^{-6}, and train for up to 20 epochs with early stopping based on a 10% validation split.

Evaluation details. Models are evaluated on Arena-Hard (Li et al., 2024), a robust instruction-following benchmark that has a 98.6% correlation with human preference. Evaluations are conducted using GPT-4o as the judge (the original benchmark used GPT-4 Turbo, which has since been deprecated). We report both raw and style-controlled (SC) win rates.