ThoughtTrace

A NEW DATA MODALITY FOR HUMAN–AI INTERACTION

ThoughtTrace

Understanding User Thoughts in Real-World LLM Interactions

Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li
Shayne Longpre, Hongxiang Gu, Maximillian Chen, Tianmin Shu

Johns Hopkins University · MIT · Google Research

THE GAP

Existing datasets record what people say.

WHAT WE CAPTURE

USER
Help me plan dinners for the week — I'm at the office four days.

ASSISTANT
Happy to help! How many people, and any dietary limits?

The visible transcript — all that prior datasets log:
WildChat (Zhao et al., 2024) **LMSYS-Chat-1M** (Zheng et al., 2023) **SWE-Chat** (Baumann et al., 2026)

THE GAP

Existing datasets record what people say. Not what they think.

WHAT WE CAPTURE

USER
Help me plan dinners for the week — I'm at the office four days.

ASSISTANT
Happy to help! How many people, and any dietary limits?

The visible transcript — all that prior datasets log:
WildChat (Zhao et al., 2024) **LMSYS-Chat-1M** (Zheng et al., 2023) **SWE-Chat** (Baumann et al., 2026)

vs

WHAT WE MISS

REASON **why they asked**

"A vague first prompt, just to get the ball rolling."

REACTION **how they felt**

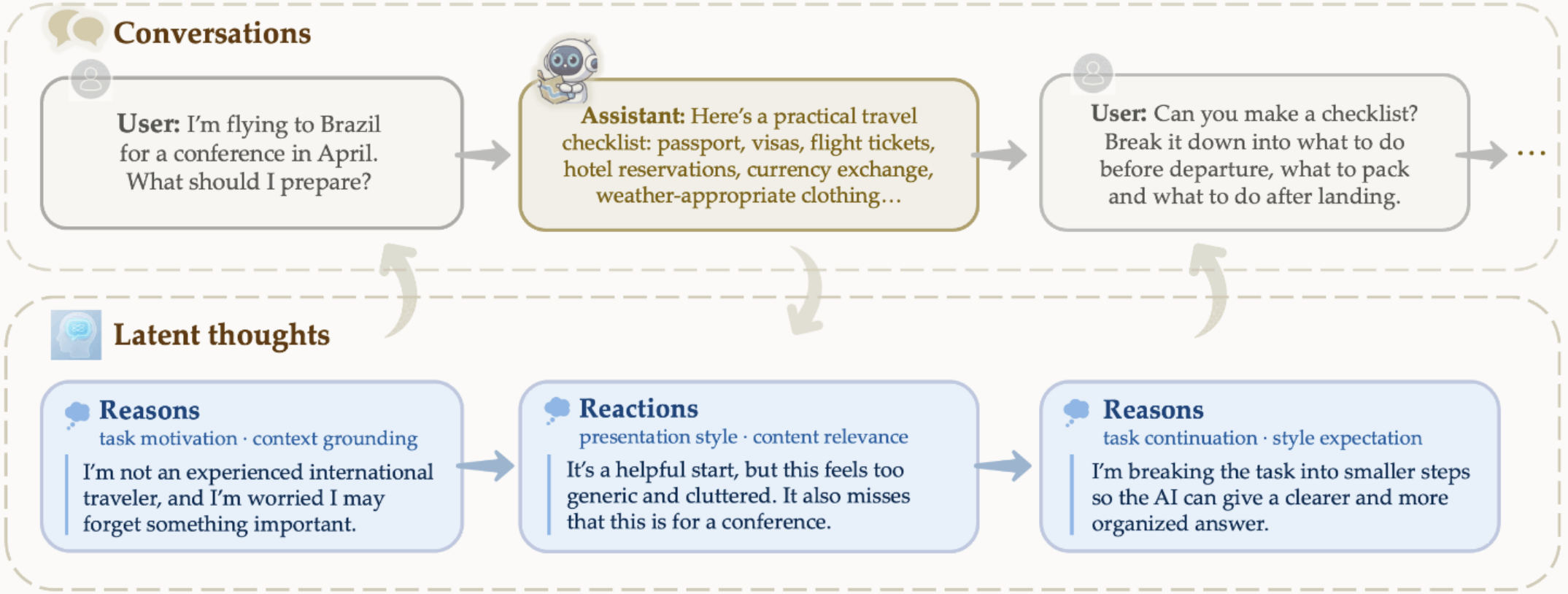
"Good — it asked for info instead of dumping suggestions."

OUR CONTRIBUTION

Introducing ThoughtTrace.

The first dataset pairing real conversations with the thoughts behind them — paper, data, and code all public.

A thought is the why behind a prompt and the reaction to a reply.



THE DATASET

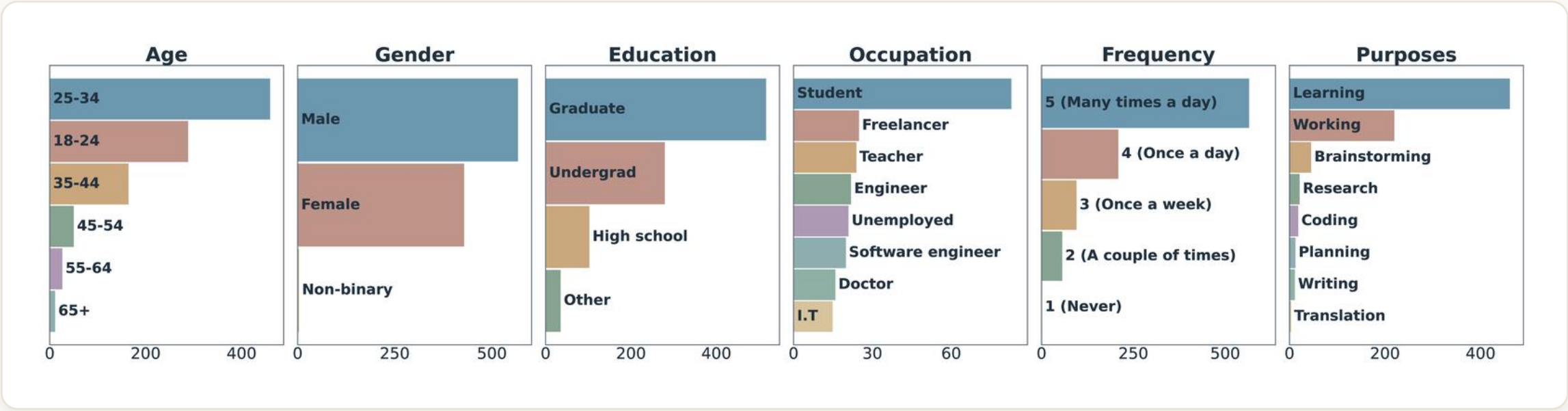
Real conversations, paired with users' own thoughts.

1,058	2,155	17,058	10,174	20
Users	Conversations	Turns	Thoughts	LLMs

Self-reported, not inferred. Real users engaged in multi-turn conversations with 20 models and annotated their own reasons and reactions throughout the interaction.

USER DEMOGRAPHICS

ThoughtTrace captures a representative spectrum of users.

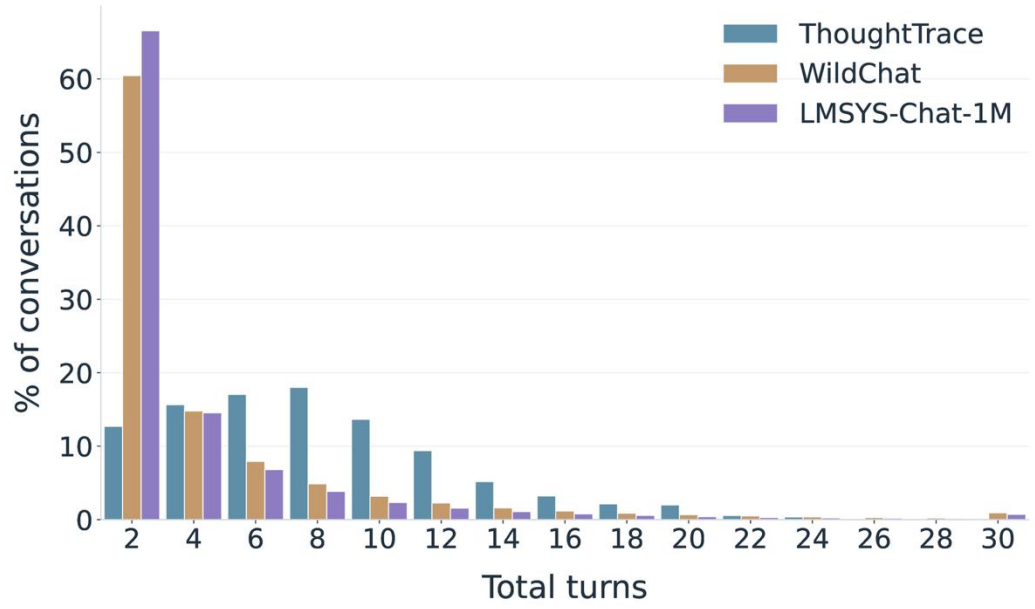


Rich demographics & usage metadata.

DATA QUALITY

Long, evolving conversations — where thoughts matter most.

Conversation length vs. prior datasets



8

median turns — vs. 2 in WildChat & LMSYS

7 · 36

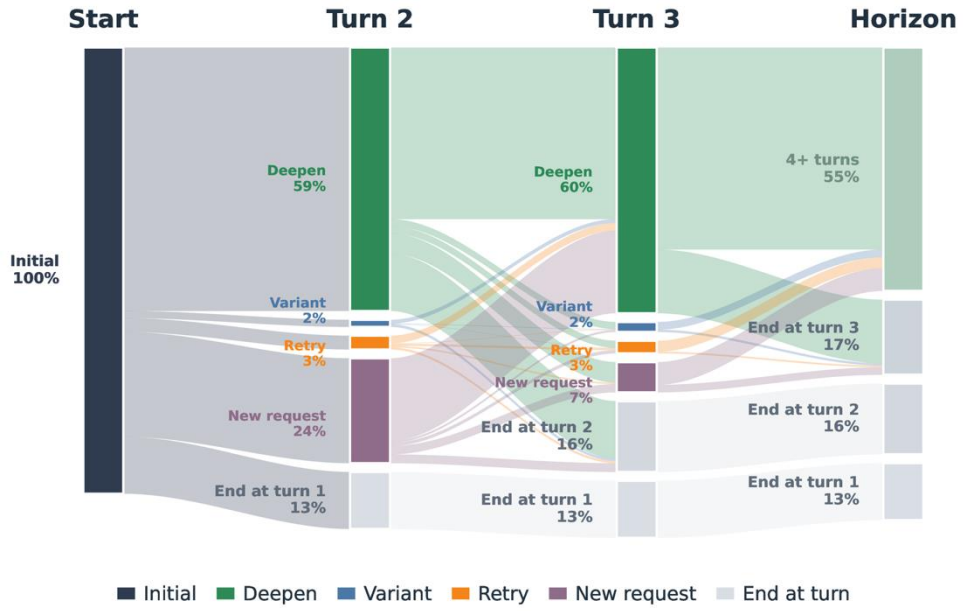
topics & subtopics — no single domain dominates

57%

of user turns — extend or build on the prior task

Turn after turn, conversations deepen rather than reset.

Turn-to-turn transitions of relationship labels



~60%

Deepen — each new turn continues the current task

55%

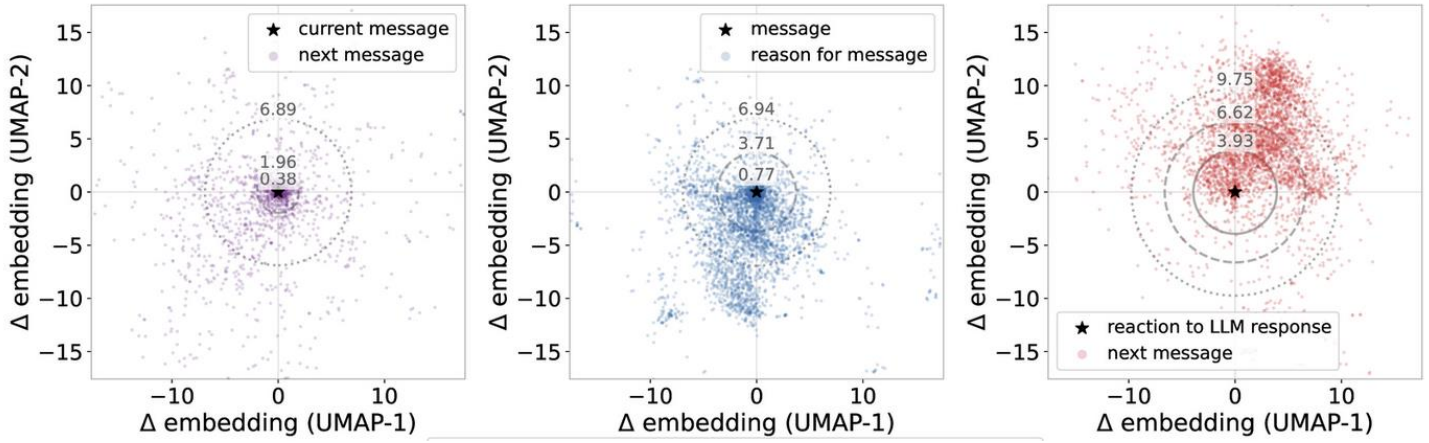
reach 4+ turns — most conversations keep going

24% → 7%

new requests fall — users settle into one task (turn 2→3)

Distinct — thoughts add what the transcript can't.

Δ embedding (UMAP) — next messages cluster tight to the current message; reasons and reactions drift much farther away.

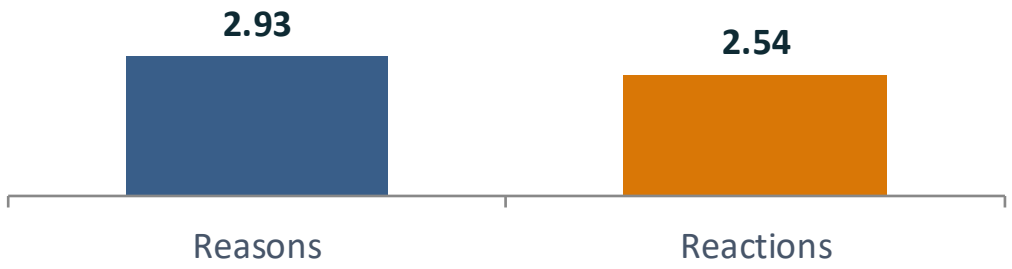


Takeaway: *thoughts carry signal you can't read off the transcript.*

THOUGHT PROPERTY · 02

Hard to infer — even frontier models can't guess them.

Recovering held-out thoughts



1–5 LLM-judge similarity. Perfect match = 5.

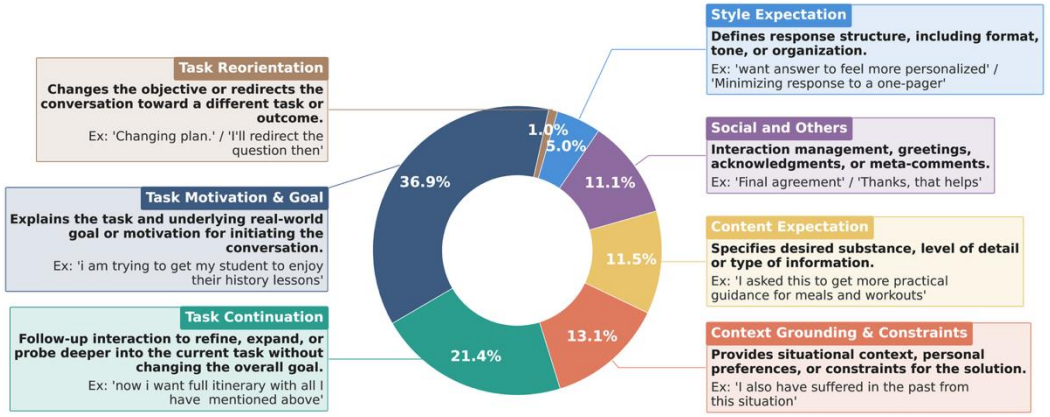
Inference is no substitute for asking.

GPT, Gemini, and Claude all struggle to infer thoughts from context—mean similarity 2.93 for reasons, 2.54 for reactions on a 1–5 scale.

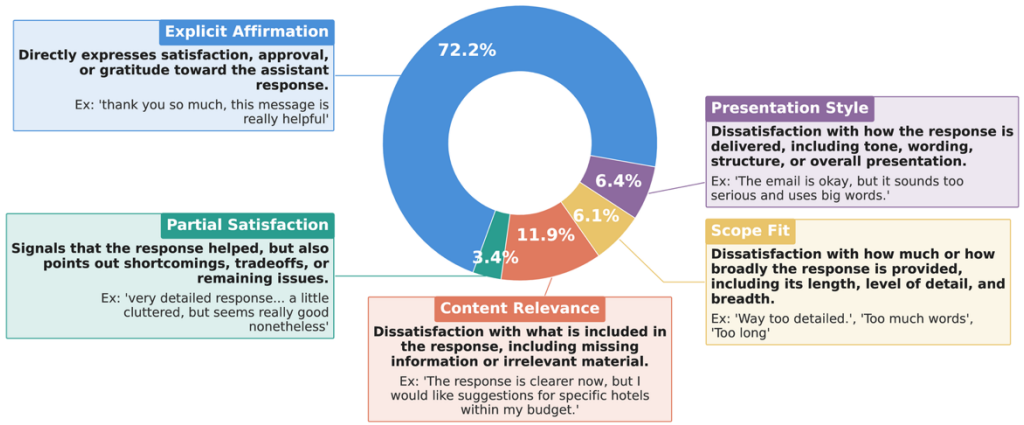
Distinct *and* **unguessable** *— together, that's why ThoughtTrace has to be collected, not derived.*

Structured & diverse — thoughts fall into clear, recurring types.

Reasons — why they ask · 7 types



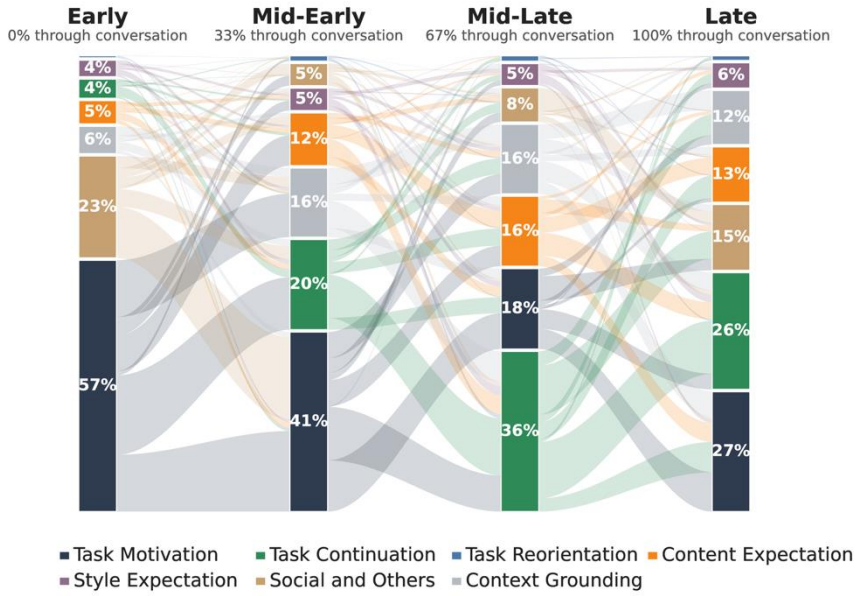
Reactions — how they feel · 5 types



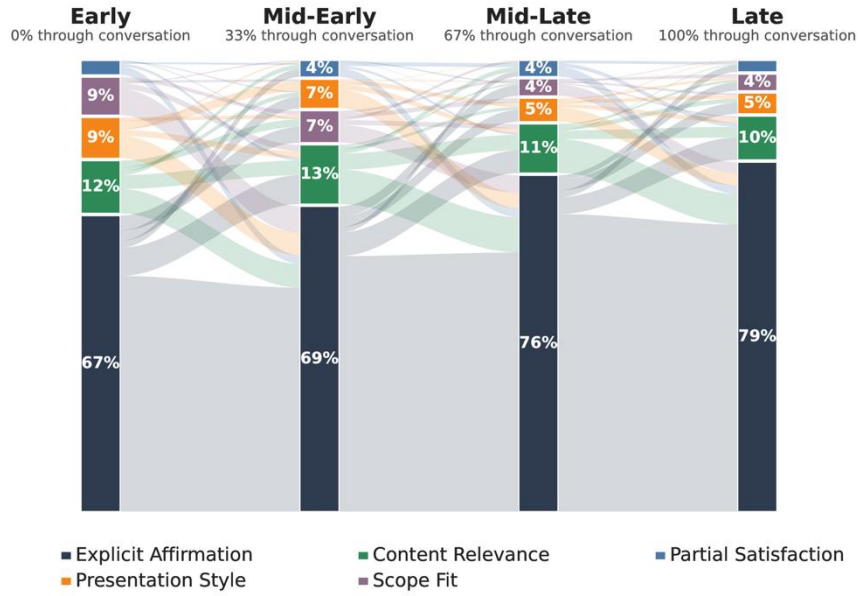
Task Motivation & Goal leads the reasons (36.9%); **Explicit Affirmation** dominates reactions (72.2%), with dissatisfaction split across content, style, and scope.

Stage-dependent — thoughts shift as the chat unfolds.

Reasons across the conversation



Reactions across the conversation



Reasons drift from motivation to continuation; **reactions** grow more affirming (67% → 79%).

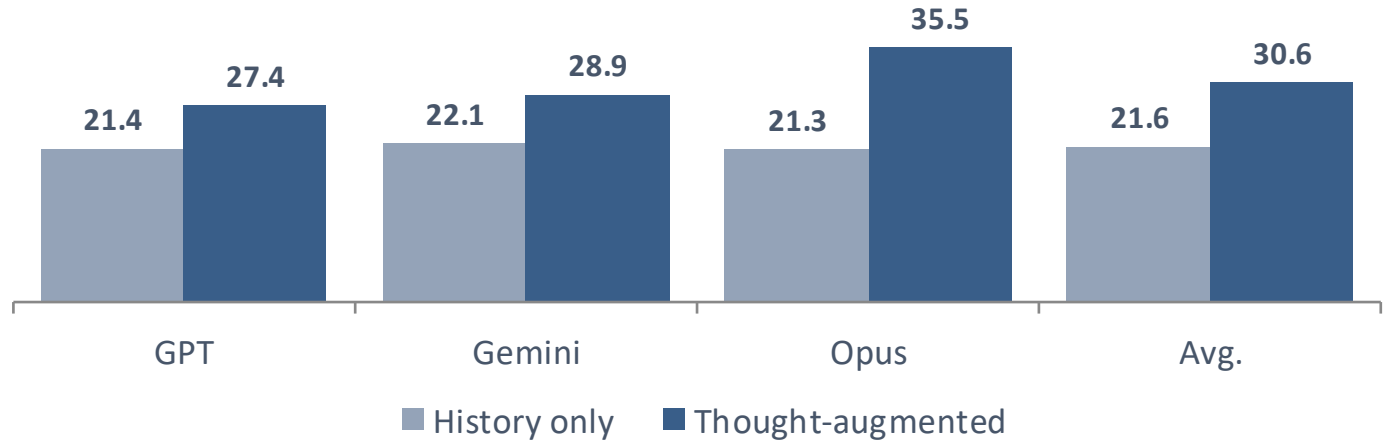
WHY IT MATTERS · 01

Thoughts predict what the user does next.

+41.7%

relative gain in next-message prediction

21.6 → **30.6**



Given the user's thoughts at inference time, three frontier models predict the next message far.

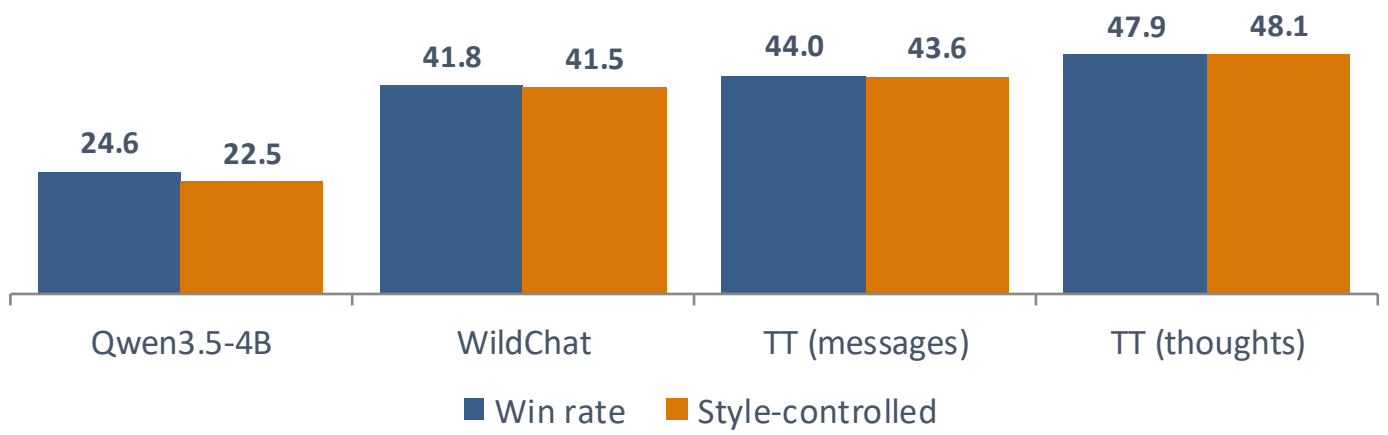
WHY IT MATTERS · 02

Thoughts are a better training signal for alignment.

+25.6%

style-controlled win rate over the base model

+4.5% vs. message-guided



Thought-guided rewrites beat message-guided rewrites, the WildChat baseline (+6.6%), and the base model — thoughts capture richer dissatisfaction and revision signal than users state out loud.

1

The first large-scale pairing of real conversations with users' self-reported thoughts.

reasons & reactions · 1,058 users · 20 models

2

1

The first large-scale pairing of real conversations with users' self-reported thoughts.

reasons & reactions · 1,058 users · 20 models

2

Thoughts are distinct, hard to infer, structured, and stage-dependent.

four properties of a genuinely complementary modality

1. **The first large-scale pairing of real conversations with users' self-reported thoughts.**

   reasons & reactions · 1,058 users · 20 models

2. **Thoughts are distinct, hard to infer, structured, and stage-dependent.**

   four properties of a genuinely complementary modality

3. **And they're actionable.**

   +41.7% behavior prediction · +25.6% alignment

TAKEAWAYS

1. **The first large-scale pairing of real conversations with users' self-reported thoughts.**

   reasons & reactions · 1,058 users · 20 models

2. **Thoughts are distinct, hard to infer, structured, and stage-dependent.**

   four properties of a genuinely complementary modality

3. **And they're actionable.**

   +41.7% behavior prediction · +25.6% alignment

WHERE IT GOES NEXT

User modeling

what users think, and how context shapes it

Model training

thoughts as a new supervisory signal to learn from

Evaluation

thought-centered measures of user satisfaction

ThoughtTrace

Thank you

ThoughtTrace — user thoughts as a new data modality for human–AI interaction.

Paper

arxiv.org/abs/2605.20087

Data

huggingface.co/datasets/SCAI-JHU/ThoughtTrace

Code

github.com/thoughttrace-project/ThoughtTrace