Learning from Real-World Human Interaction

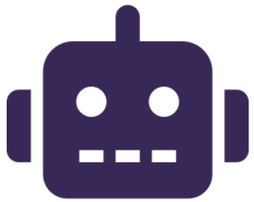
What users do, and what they think

Chuanyang Jin

Johns Hopkins University

THE SETTING

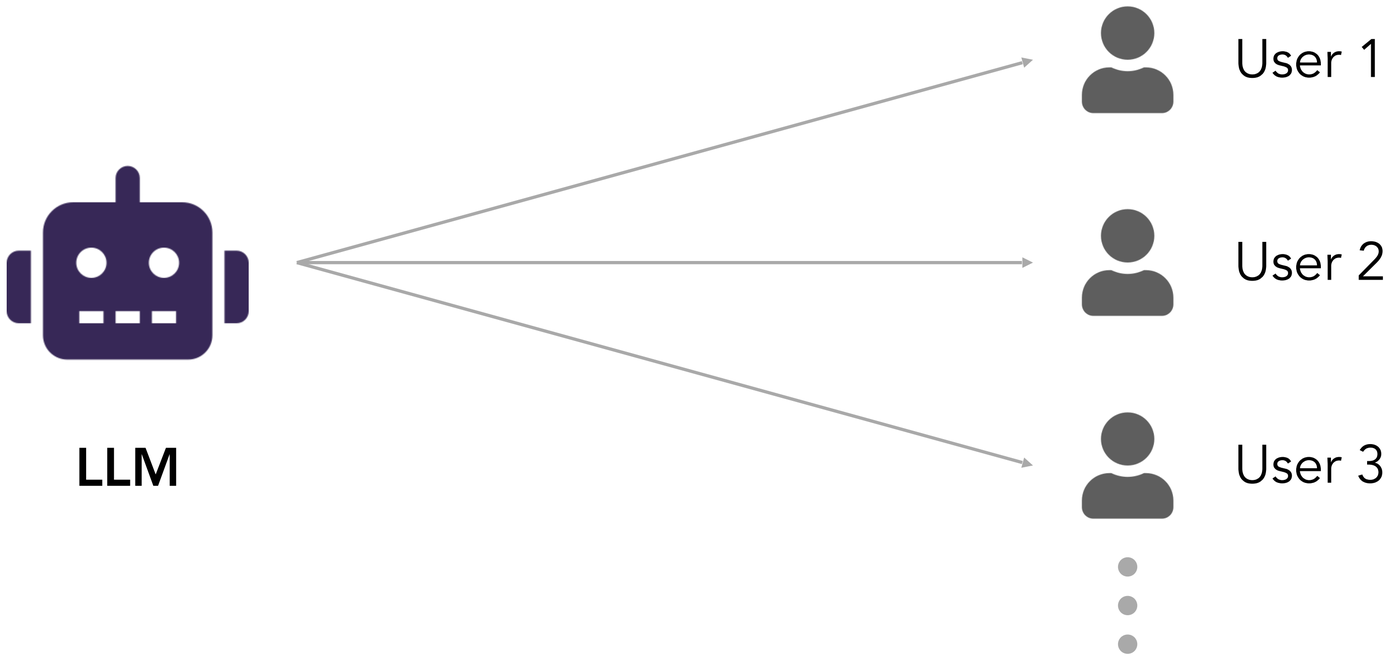
LLMs are deployed to interact with **users**.



LLM

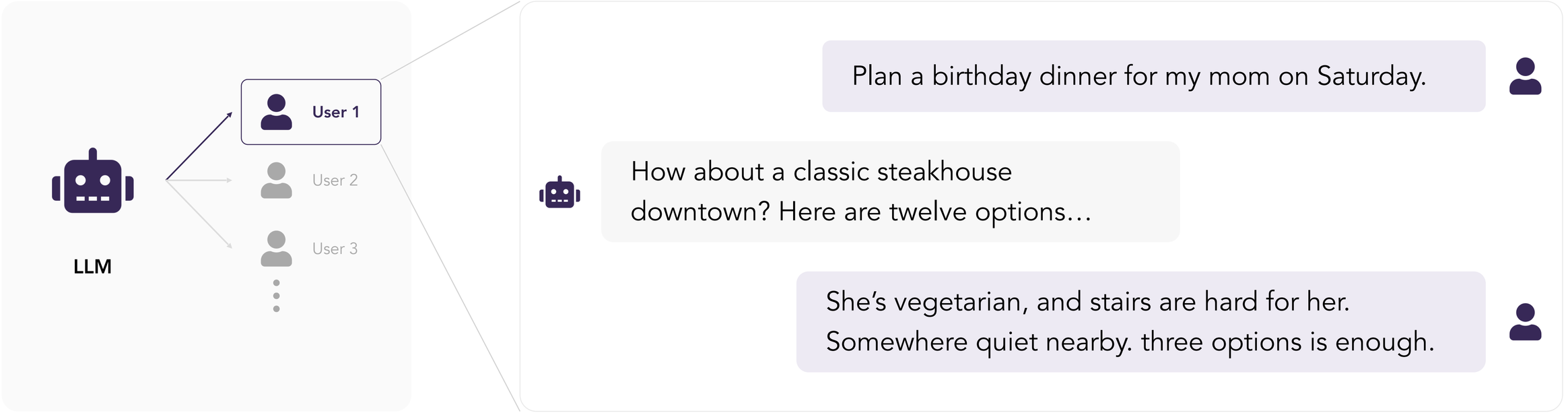
THE SETTING

LLMs are deployed to interact with **users**.



ONE INTERACTION

Zoom in on **a single user**.



WHAT THE MODEL CAN LEARN

One conversation, three kinds of **supervision**.

Knowledge

Mom is vegetarian and can’t manage stairs — facts no dataset holds.

Skills

Ask about constraints before recommending — a better way to help.

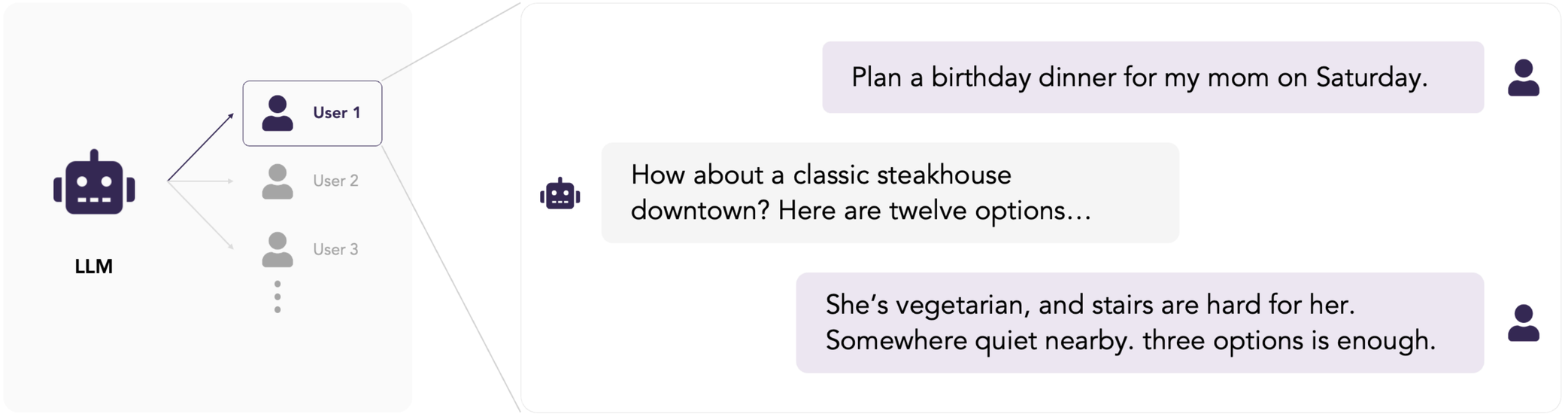
Personal preferences

Quiet, close to home, three options — not twelve.

…and more: context, tone, values. **Every turn is a signal.**

THE VISION

We advocate for an era of **real-world human interaction**, where models learn from natural human interaction for continual improvement and personalized alignment.

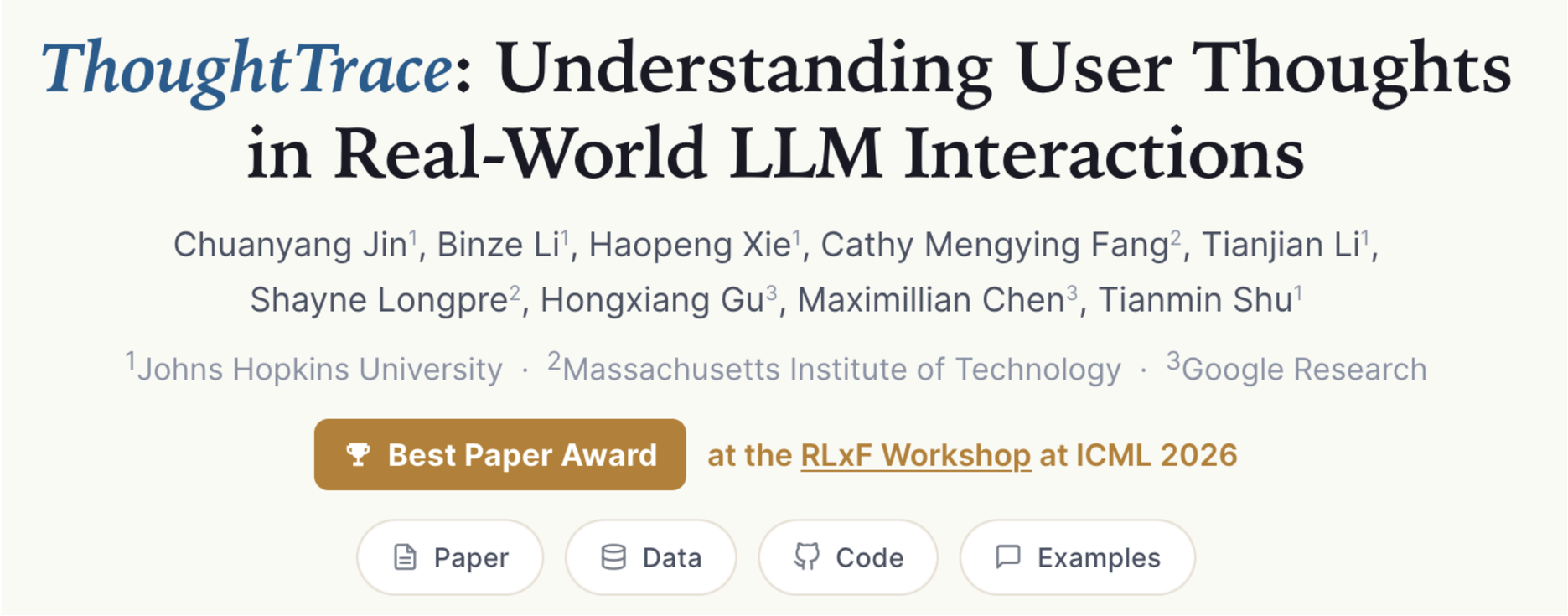


We introduce **Reinforcement Learning from Human Interaction (RLHI)** as a first step toward this vision.

RLHI · Jin et al., 2025

THE NEXT STEP

We present **ThoughtTrace**, introducing users' unspoken thoughts as a new data modality for understanding real users and improving models.



ThoughtTrace · Jin et al., ICML 2026 RLxF Best Paper Award

Two steps toward learning from real human interaction

PART I

RLHI

Jin et al., 2025

Learn directly from user conversations — organic replies and long-term history become the training signal.

PART II

ThoughtTrace

Jin et al., ICML 2026 RLxF · Best Paper Award

Capture the reasons and reactions behind each turn — a modality no transcript records.

PART I

Reinforcement Learning from Human Interaction

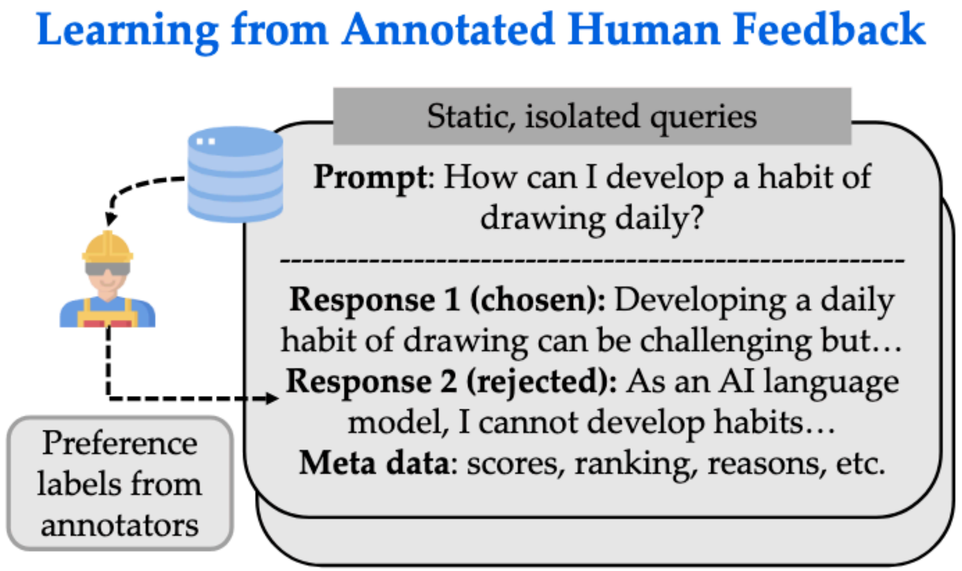
Learning directly from real user conversations

JOHNS HOPKINS UNIVERSITY · META FAIR

THE PROBLEM

Post-training still learns from **annotators/fixed tasks**, not from **users**.

Verifiable tasks, fixed demonstrations, and ratings collected outside natural conversations.



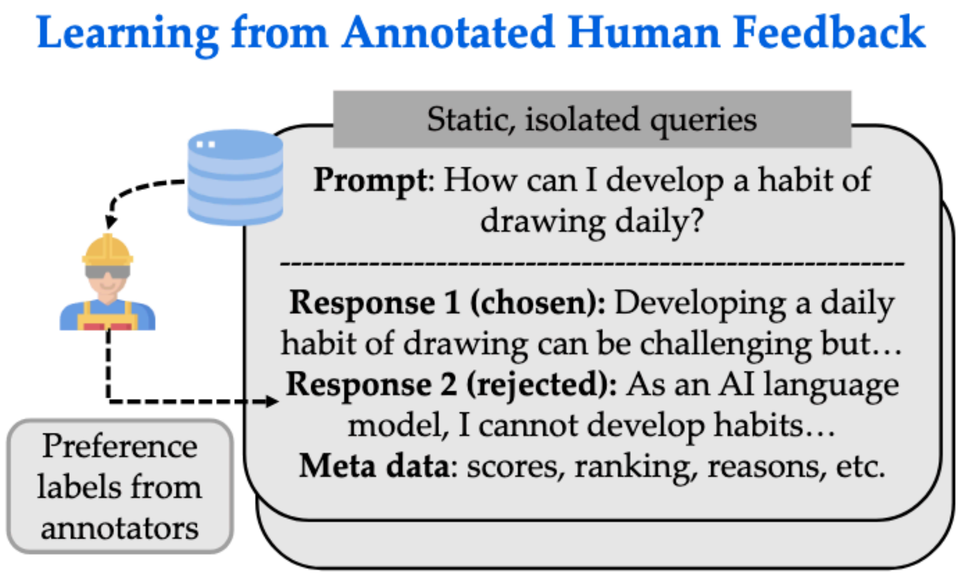
THE PROBLEM

From **annotated feedback** to **organic human interaction**.

Annotator opinions and heuristics

Static, context-free judgments

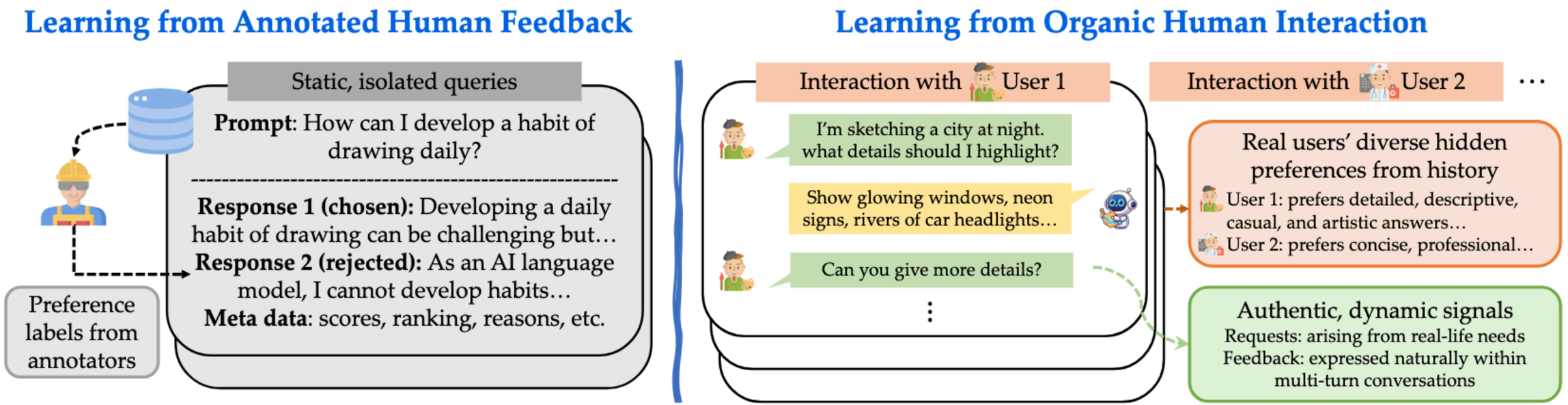
Scaling with labeling budgets



THE SHIFT

From **annotated feedback** to **organic human interaction**.

Annotator opinions and heuristics	vs.	**Real users' diverse goals and preferences**
Static, context-free judgments	vs.	**Evolving, situational demands**
Scaling with labeling budgets	vs.	**Scaling with actual usage**



AI advances by learning continually from **interaction** — and human interaction is its core pillar.

Contextual grounding

Tied to users' situational needs, the model's prior outputs, and personal history.

Evolving distribution

Reflects shifting goals and environments over time, not a frozen snapshot.

Diverse supervision

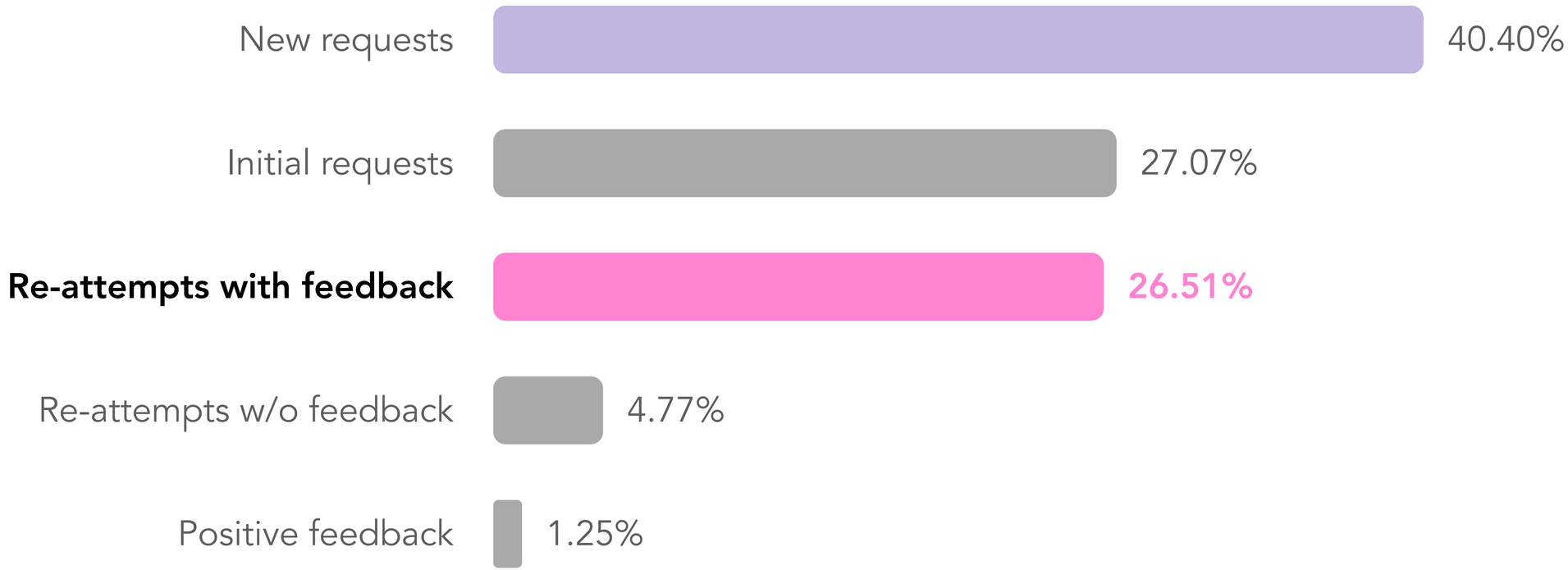
Explicit corrections and clarifications, plus implicit cues like tone and disengagement.

In its minimal form the substrate already exists: LLMs converse with **millions of users** every day.

Silver & Sutton, "Welcome to the Era of Experience"

FINDING 01 · WILDCHAT-1M

In the wild, users constantly **correct the model**.



Later turns are dominated by re-attempts with feedback — **83% after the 5th turn**.

Feedback messages are **shorter, yet semantically dense**.

Re-attempt With Feedback – Example 1

User: give me 20 ideas for themes for a summer camp for children aging from 4 to 18

Assistant: 1. Nature and Outdoor Adventure... (642 characters skipped here)

User: give me 20 more make them creative

The third turn — **"give me 20 more make them creative"** — is the whole training signal: seven words naming what to fix.

Users refine the prompt, add clarification, or react to what the model got wrong, all inside the conversation.

Real conversations are more **diverse** — and users want **different things**.

Dimension	Preference 1	Pct.	Preference 2	Pct.	Pct. None
expertise	responses that can be easily understood by beginners	24.1%	responses with expert-level knowledge	**59.8%**	16.1%
informativeness	concise responses, without being verbose	36.0%	expansive and informative responses, without missing background information	**49.9%**	14.1%
tone	casual, friendly, and humorous responses	4.9%	serious, formal, and professional responses	**84.5%**	10.6%
structure	structured responses, with a clear and logical flow	**77.1%**	free-form responses, with a casual and conversational style	9.1%	13.8%

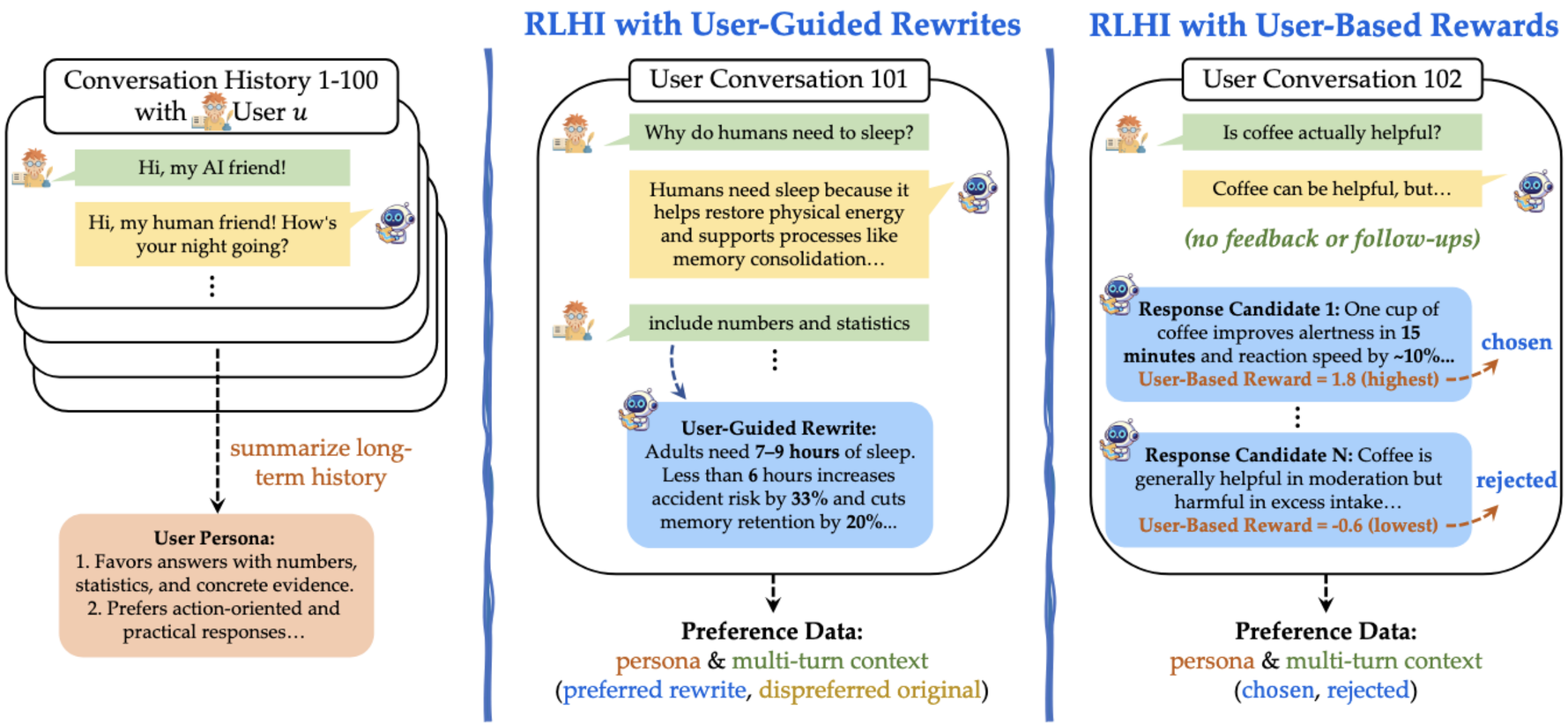
WildChat reaches **0.865** contextual diversity — above HH-RLHF and HelpSteer2.

Some personas are common; others are unique, and one user can want detailed math but concise advice.

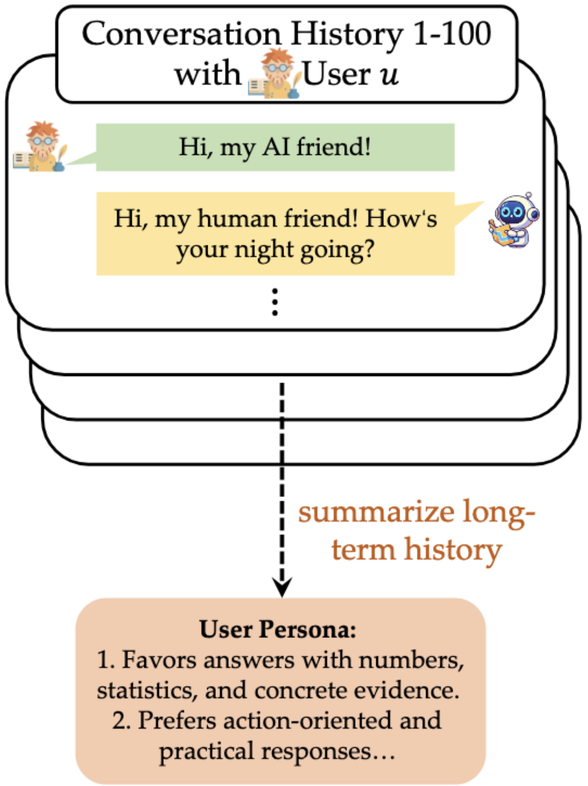
Can we learn directly from
user conversations?

THE METHOD

Reinforcement Learning from Human Interaction (RLHI) trains on **organic replies** and **long-term history**.



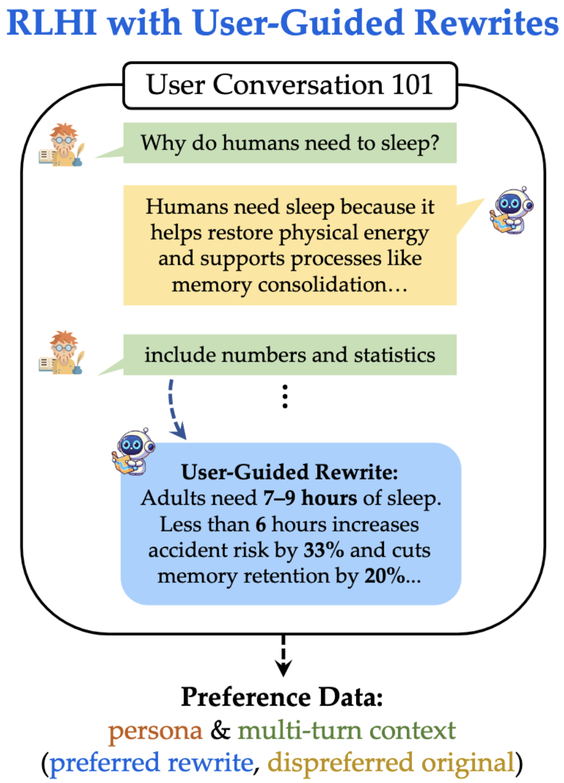
A **persona** summarizes each user's long-term history.



- Derived in natural language from all of a user's prior conversations.
- Carries stable preferences — format, tone, depth, evidence.

METHOD 01

RLHI with **User-Guided Rewrites**

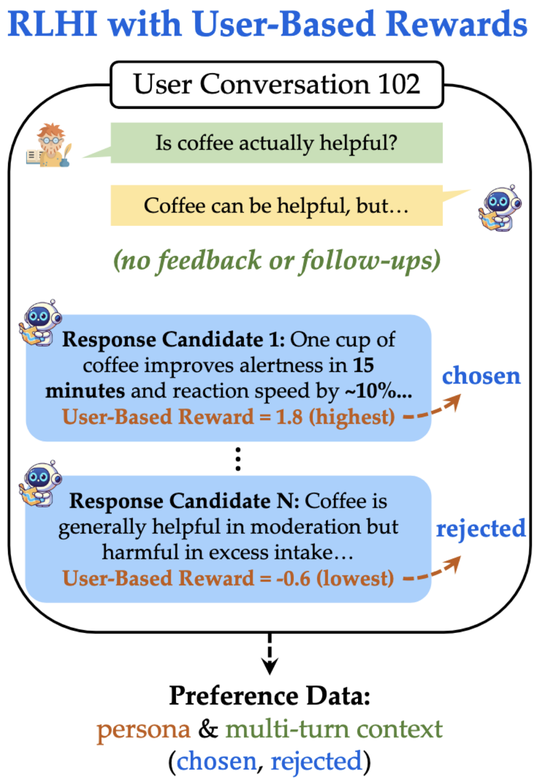


1. Find re-attempts with feedback — 26.5% of user messages.
2. Ask the model to revise its prior response using that feedback.
3. Pair the user-guided rewrite (preferred) against the original.
4. Keep only high-quality pairs.
5. Train with persona-conditioned DPO.

Organic feedback is rich text, not a binary label — so use it to locate the weakness and patch it.

METHOD 02

RLHI with **User-Based Rewards**



1. Sample N candidate responses for a real user request.
2. Score each with a reward model conditioned on the user persona.
3. Pair the highest- against the lowest-scoring candidate.
4. Train with persona-conditioned DPO.

Most requests arrive without feedback — yet they still reflect genuine needs, so this variant can run **online**.

THE OBJECTIVE

Both variants optimize **persona-conditioned DPO**.

$$\mathcal{L}_{\text{persona-DPO}} = \mathbb{E}_{u,i}\left[\log \sigma\Big(\beta\Big(\log \frac{\pi_\theta(y^+_{u,i}|x_{u,i},p_u)}{\pi_{\text{ref}}(y^+_{u,i}|x_{u,i},p_u)} - \log \frac{\pi_\theta(y^-_{u,i}|x_{u,i},p_u)}{\pi_{\text{ref}}(y^-_{u,i}|x_{u,i},p_u)}\Big)\Big)\right]$$

The persona and the multi-turn context enter the conditioning, so the same prompt can be answered differently for different users.

EVALUATION

WildChat UserEval: **real users**, **real histories**.

Personalization

Which response better aligns with the user's persona?

Instruction-Following

Which response is more faithful and higher quality?

UserEval

Holistic, user-like judgment of the whole answer.

Users with ≥ 10 conversations · first N−5 for reference history, last 5 held out.

Judged by o3 · two inference modes: Context-Only and Persona-Guided.

RLHI wins on **personalization** and **instruction-following**.

	Personalization	Instr-Following	UserEval
Baselines			
Llama-3.1-8B-Instruct	38.2	30.6	32.5
+ *Persona-Guided Inference*	39.8	29.2	31.3
RL with Rewrites from Scratch	52.5	41.3	46.3
+ *Persona-Guided Inference*	54.6	40.4	47.3
RL with User-Agnostic Rewards	52.7	43.3	47.9
+ *Persona-Guided Inference*	54.2	42.8	48.4
RLHI			
User-Guided Rewrites	54.6	45.5	52.0
+ *Persona-Guided Inference*	**62.5**	44.5	**54.9**
User-Based Rewards	61.0	**46.8**	51.3
+ *Persona-Guided Inference*	62.3	44.7	52.5

Beats the Instruct model, RLHF on top of it, and rewriting without conversational feedback — with a human study agreeing.

The gains **hold** where no user is in the picture.

	AlpacaEval2		Arena-Hard
Standard models	LC Win	Win	Score
Llama-3.1-8B-Instruct	20.9	21.8	21.3
RL with Rewrites from Scratch	34.7	31.0	50.0
RL with User-Agnostic Rewards	77.0	73.3	**64.4**
RLHI with User-Guided Rewrites	35.2	38.5	51.2
RLHI with User-Based Rewards	**77.9**	**83.4**	64.3

User-Based Rewards gain most; User-Guided Rewrites gain less — multi-turn real data against single-turn benchmarks.

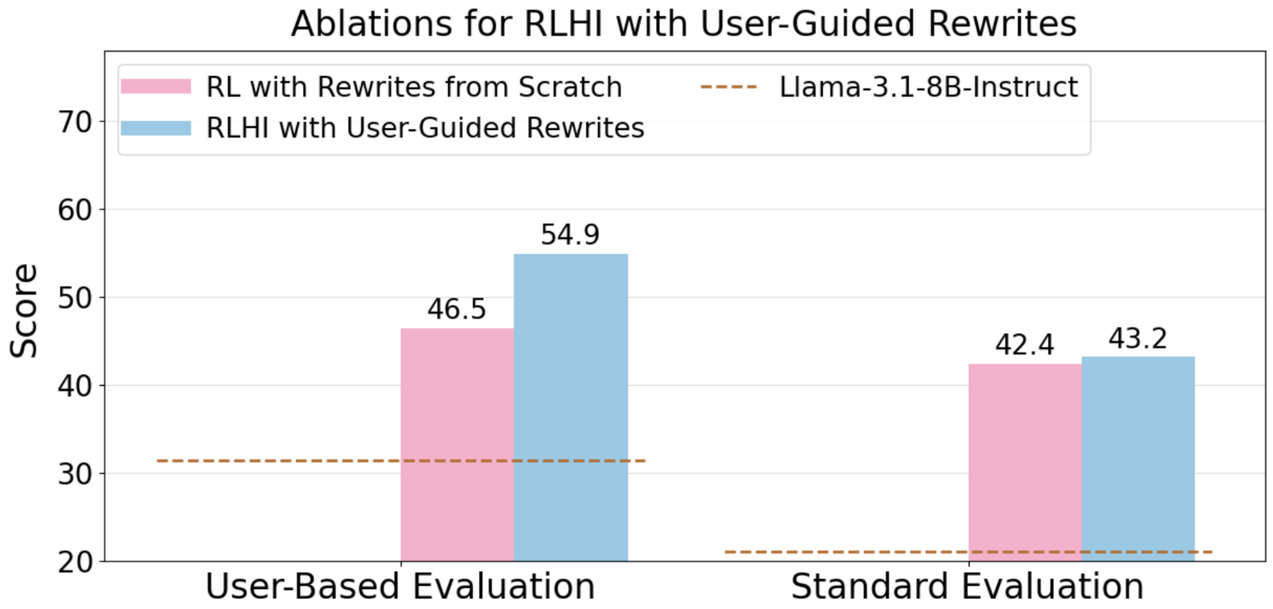
Even **vague user comments** improve reasoning.

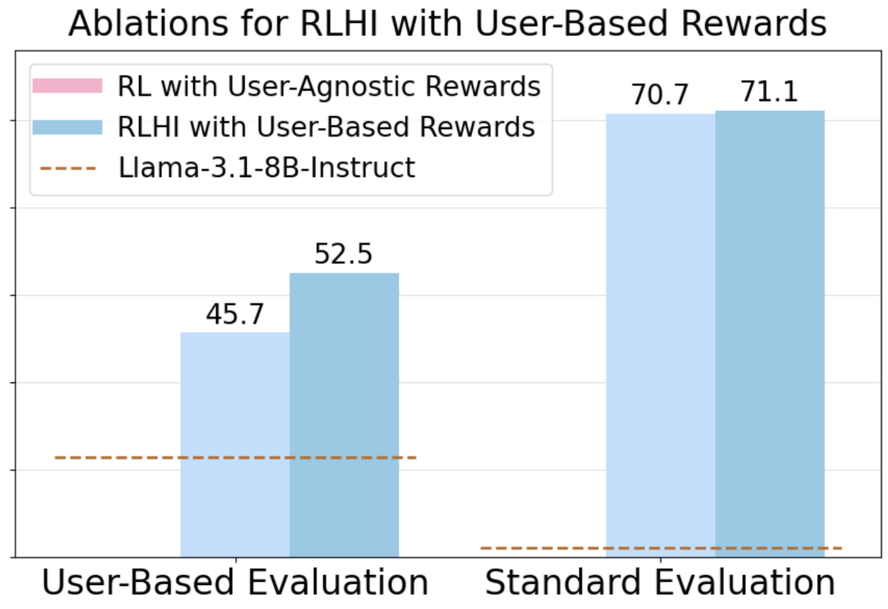
	Minerva	Olympiad	GPQA	MMLU-Pro	Avg.
Llama-3.1-8B-Instruct	20.2	14.5	26.3	44.9	26.5
RLHI with User-Guided Rewrites	**25.4**	**18.4**	**33.1**	**50.1**	**31.8**

Simulated math conversations from PRM800K — users only say "Step 3 seems incomplete", never the answer.

Average across Minerva, Olympiad, GPQA, and MMLU-Pro: **26.5 → 31.8** — the gains generalize beyond math.

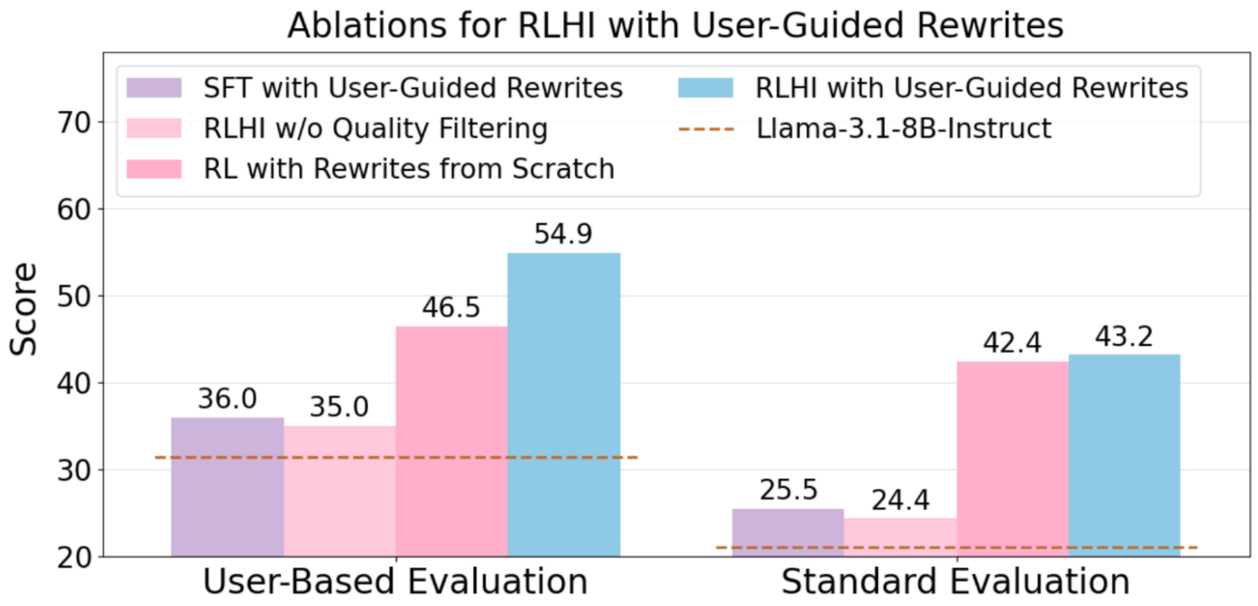
Context beats regeneration; **long-term** preference beats **generic** preference.

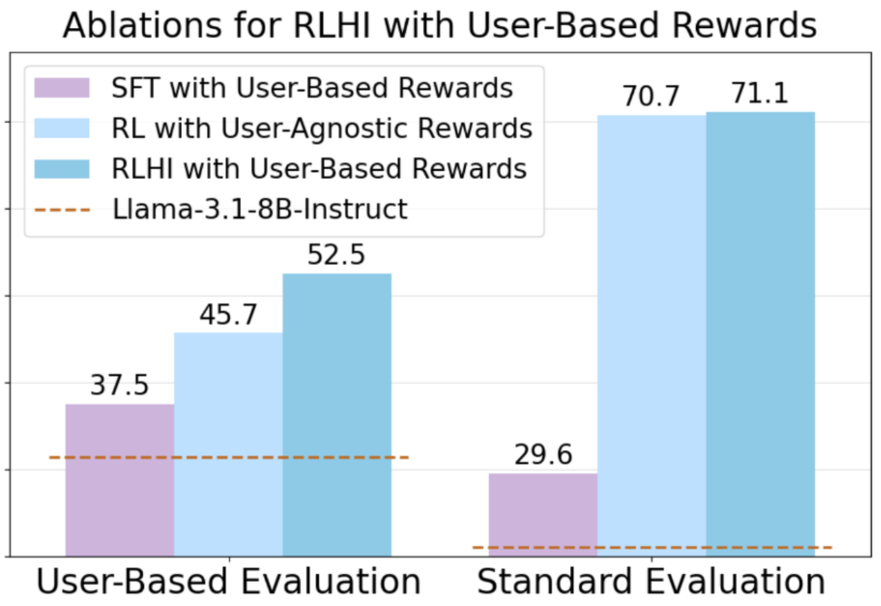




Feedback tells the model what to fix; the persona tells it what this user has always wanted.

RL extracts more from interaction than supervised fine-tuning.

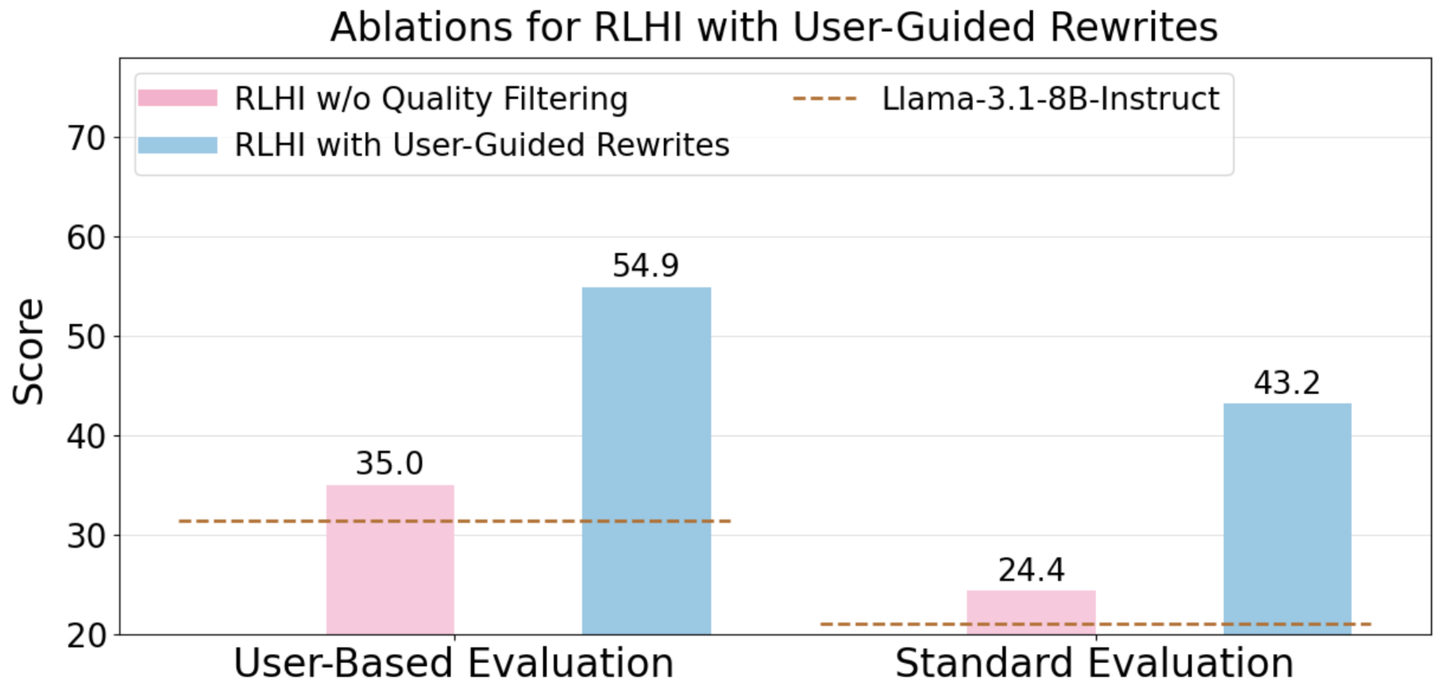




Preference optimization uses the contrast between the original and the rewrite; SFT throws that contrast away.

ANALYSIS · DATA QUALITY

Interaction data is **noisy** — filtering does real work.



Drop the filter and user-based evaluation falls **54.9 → 35.0**; standard evaluation falls **43.2 → 24.4**.

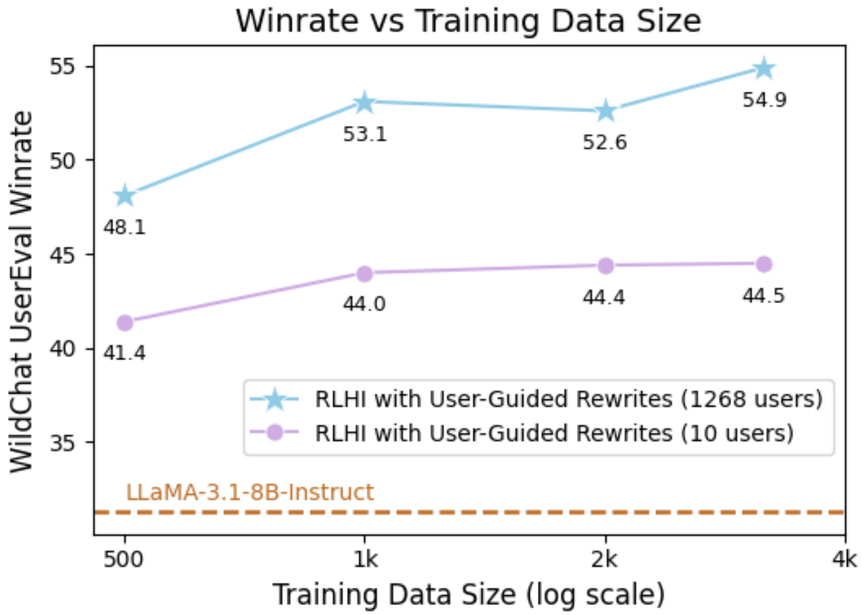
The filter is simple: keep only rewrites that improve on the original, then keep only high-quality pairs.

ANALYSIS · SCALING

RLHI scales with **user diversity**, not just data volume.

Same data budget, **126× more users**: 1,268 users reach 54.9 where 10 users stall at 44.5.

The 10-user curve flattens after 1k examples. The diverse curve keeps **climbing**.



Real-world usage plus user diversity could be the **next scaling law**.

Organic human interaction is scalable, effective supervision for personalized alignment.

RLHI is a simple, concrete method — and it works.

RLHI learns from what users **say**.
But users don't say everything.

The reason behind a message, and the reaction to a reply, are never written down.

PART II

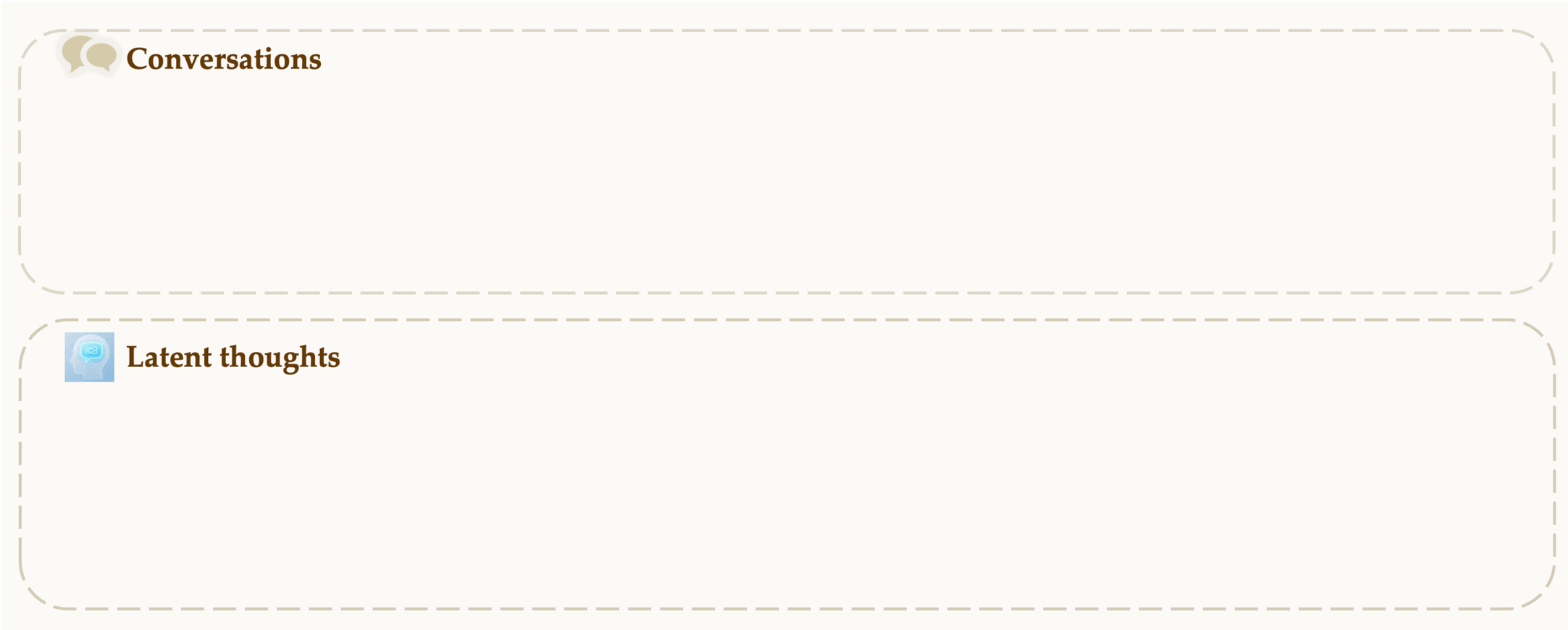
ThoughtTrace

User thoughts as a new data modality

ICML 2026 RLxF Workshop · Best Paper Award

THE GAP

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



WildChat, LMSYS-Chat-1M, and SWE-Chat all log the visible transcript. The reason and the reaction are never captured.

THE DATASET

Real conversations, paired with users' **own thoughts**.

1,058
Users

2,155
Conversations

17,058
Turns

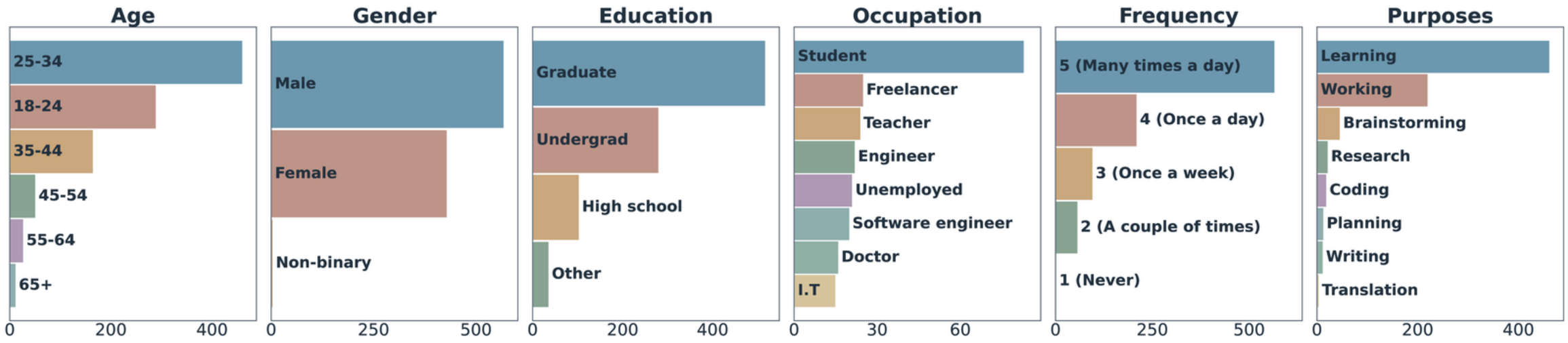
10,174
Thoughts

20
LLMs

Self-reported, not inferred. Real users held multi-turn conversations with 20 models and annotated their own reasons and reactions as they went.

USER DEMOGRAPHICS

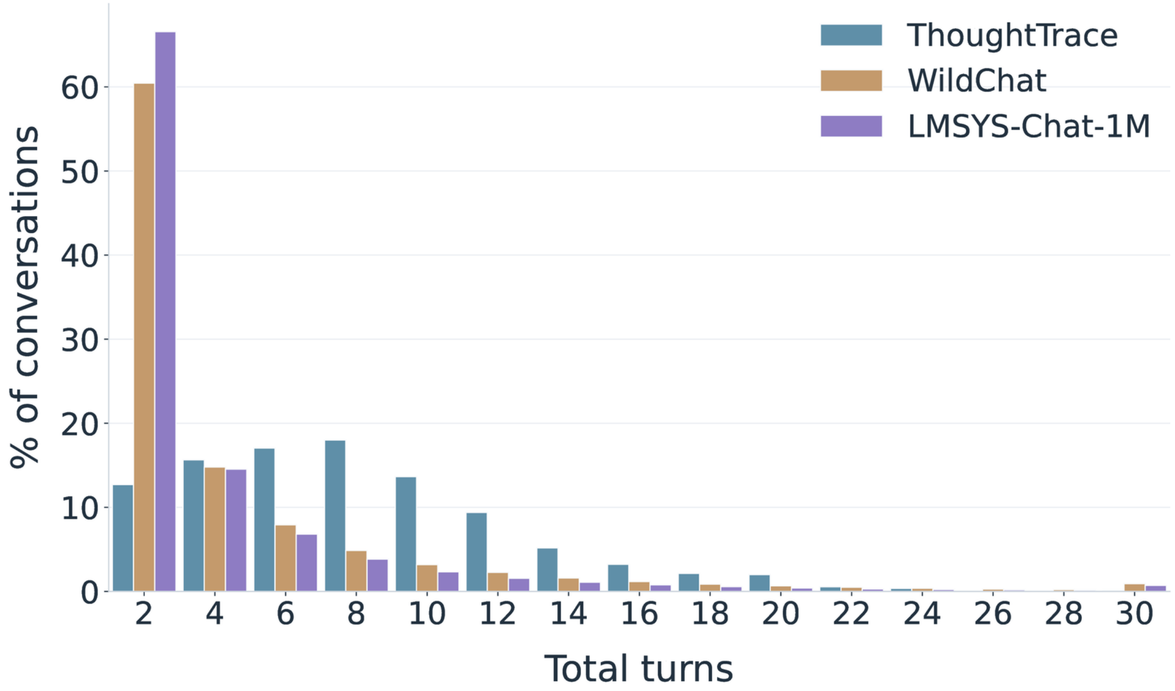
A **representative spectrum** of real users.



Rich demographics and usage metadata for every participant.

DATA QUALITY

Long, evolving conversations — where thoughts **matter most**.



Conversation length vs. prior datasets

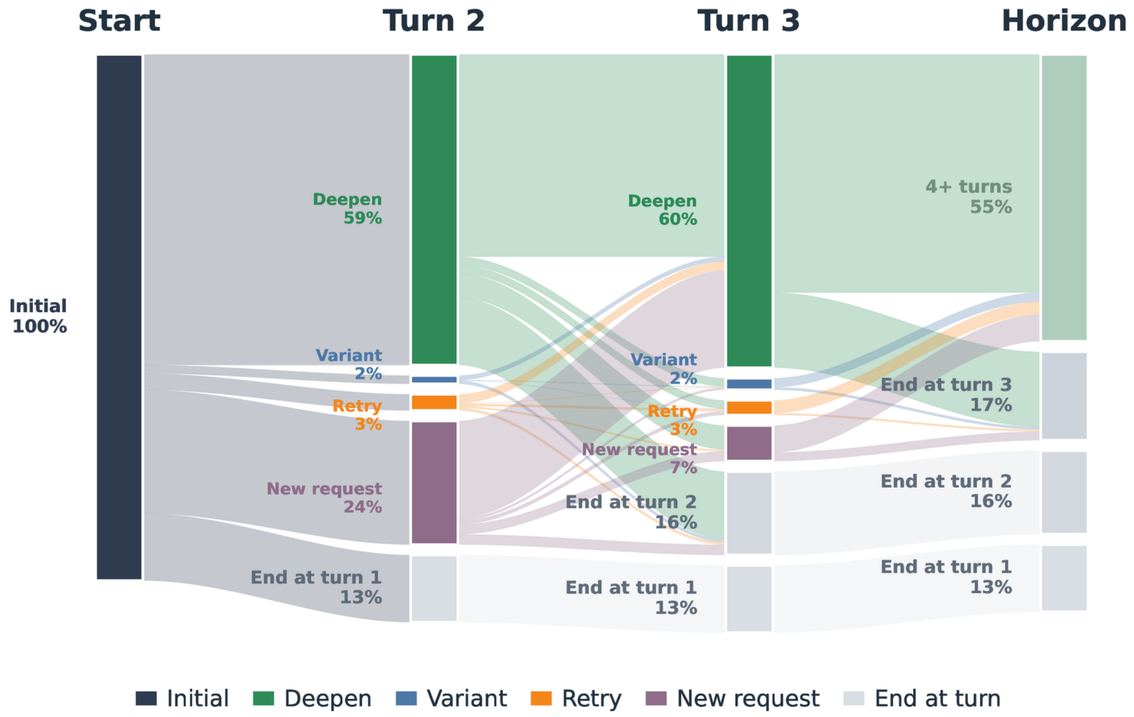
8 median turns — vs. 2 in WildChat and LMSYS

7 · 36 topics and subtopics — no single domain dominates

57% of user turns extend or build on the prior task

TURN-TO-TURN DYNAMICS

Turn after turn, conversations **deepen** rather than reset.



Turn-to-turn transitions of relationship labels

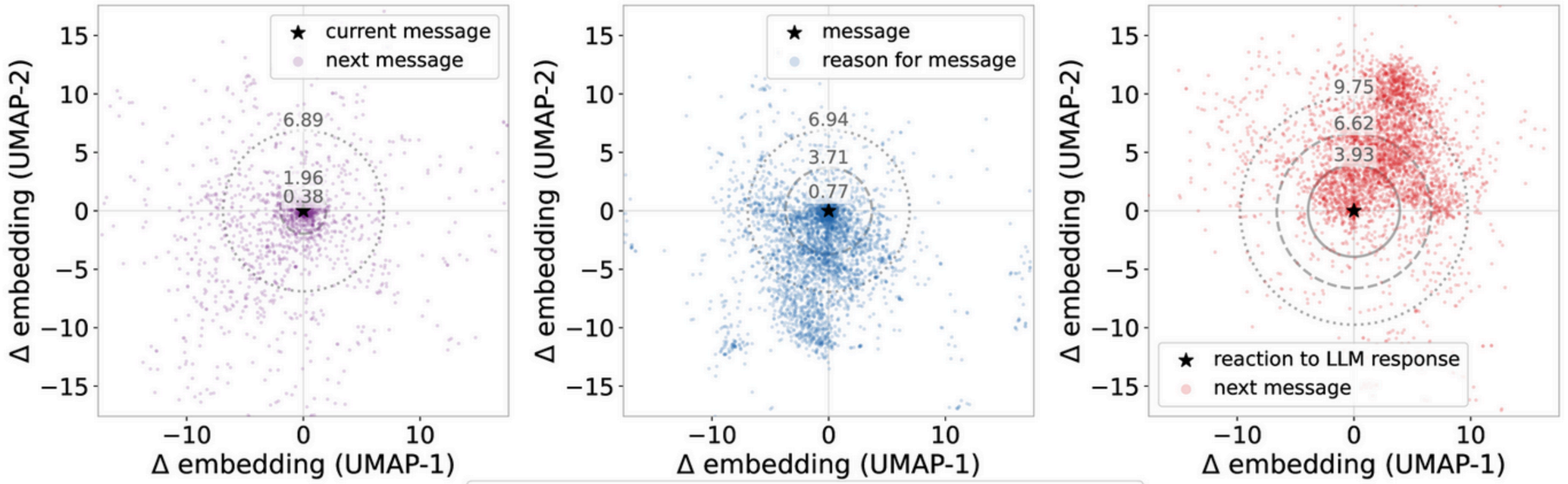
~60% deepen — each new turn continues the current task

55% reach 4+ turns — most conversations keep going

24% → 7% new requests fall as users settle in (turn 2→3)

THOUGHT PROPERTY · 01

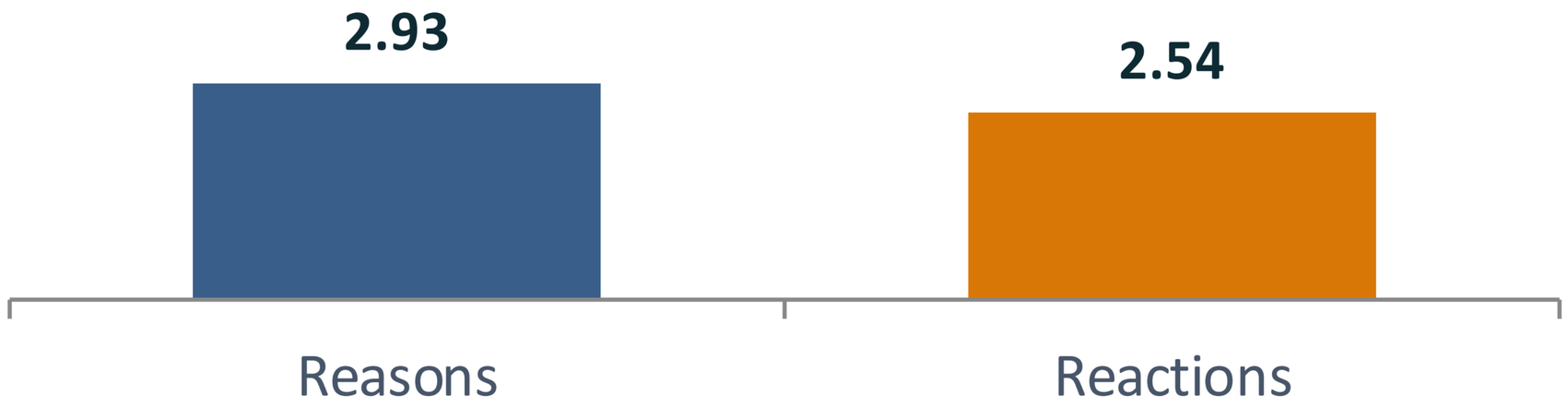
Distinct — thoughts add what the transcript can't.



Next messages cluster tight to the current message; reasons and reactions drift far away — the transcript has neither.

THOUGHT PROPERTY · 02

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

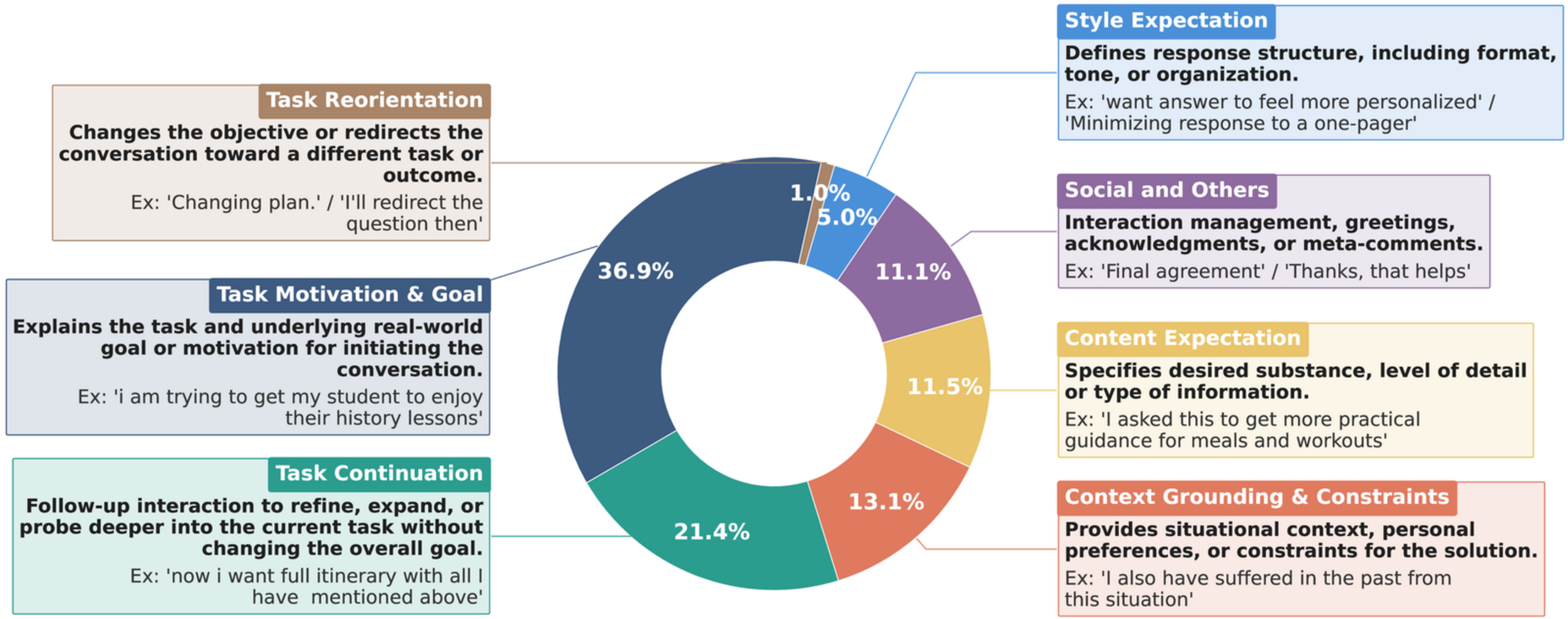


Recovering held-out thoughts · 1–5 LLM-judge similarity, perfect match = 5

Even the strongest models land mid-scale — **2.93** for reasons and **2.54** for reactions. Inference is no substitute for asking.

THOUGHT PROPERTY · 03

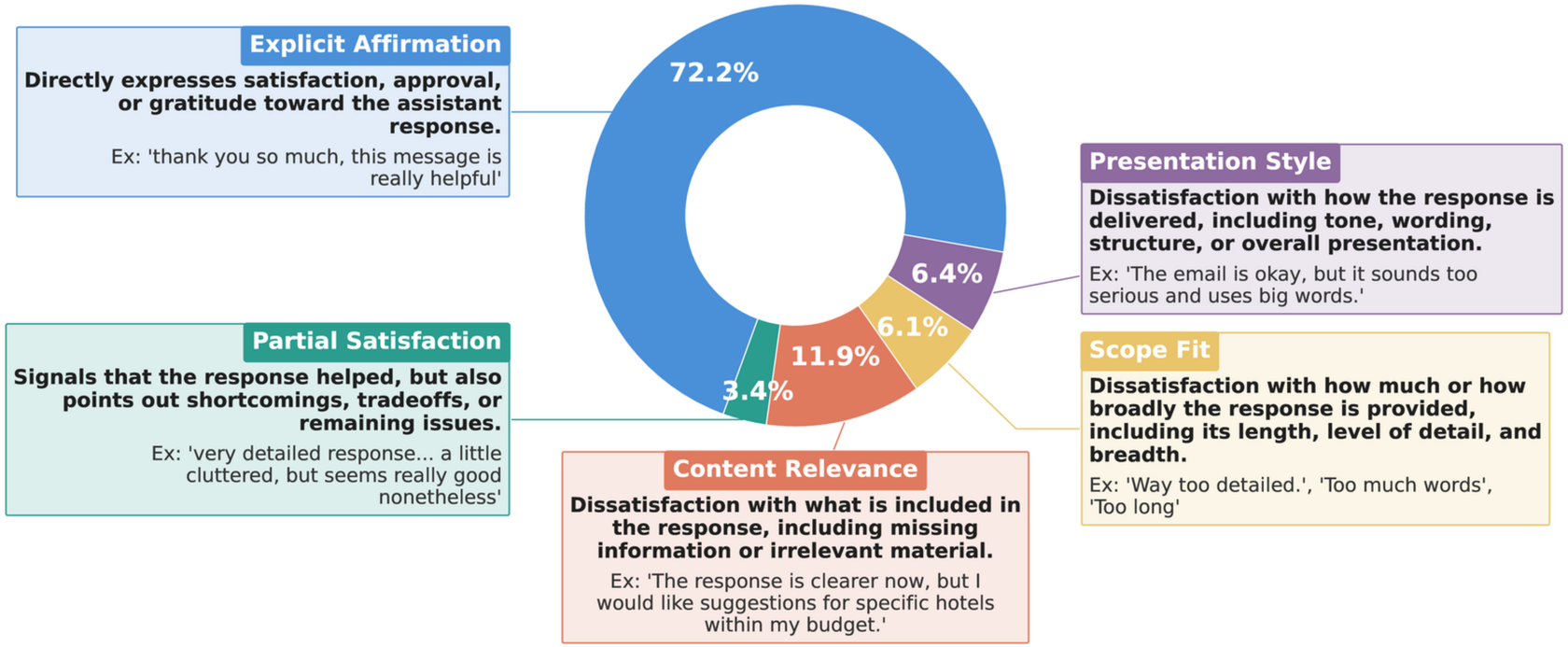
Structured — seven recurring reasons behind a message.



Task Motivation & Goal leads (**36.9%**) — users mostly explain the real-world goal a transcript never states.

THOUGHT PROPERTY · 03

...and five recurring **reactions** to a reply.

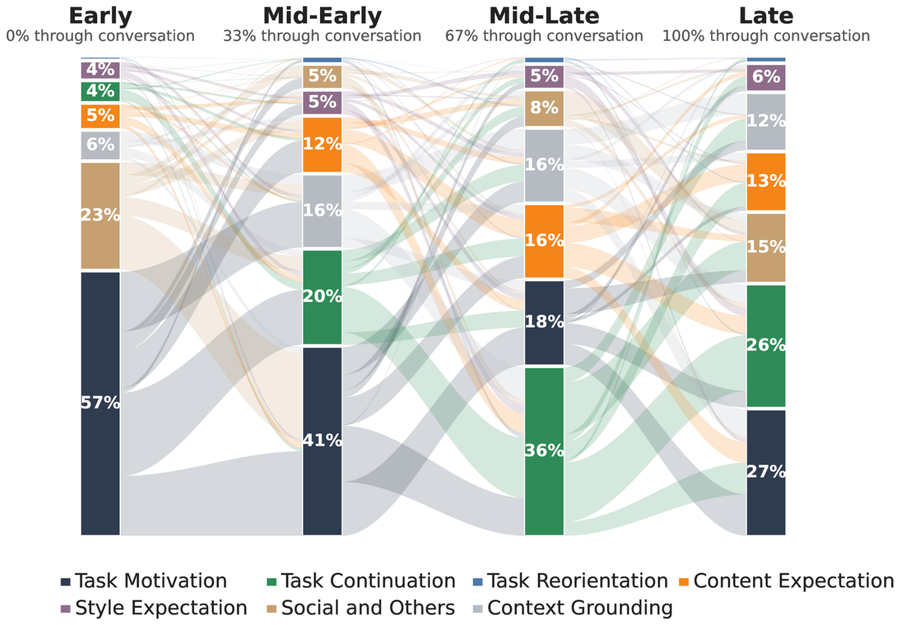


Explicit Affirmation dominates (**72.2%**) — but the remaining 28% splits dissatisfaction across content, style, and scope.

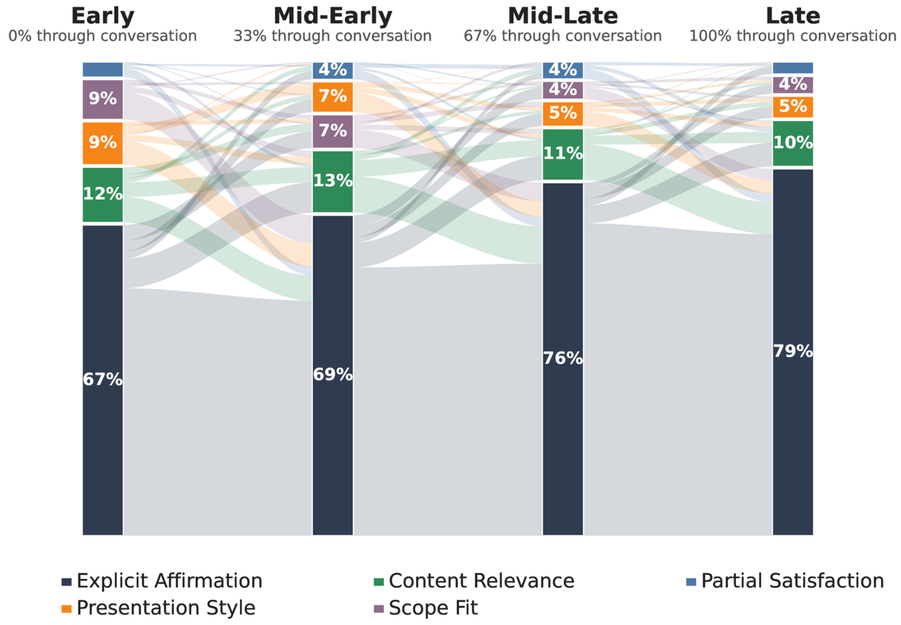
THOUGHT PROPERTY · 04

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 · THE SETUP

User-behavior prediction: we interleave the user's **thoughts** into the transcript, then ask for the next message.

User: I'm flying to Brazil for a conference in April. What should I prepare?

Note: The user's reason for sending this message is: worried about forgetting something important.

Assistant: Here is a practical travel checklist: passport, visas, flights, hotel reservations…

Note: The user's reactions to this message are: too generic and cluttered · misses the conference.

→ **Predict:** Can you make a checklist — before departure, what to pack, and after landing?

The two **Note:** lines are the only added inputs — a transcript already contains everything else.

Scored by semantic similarity to the message the user actually sent next.

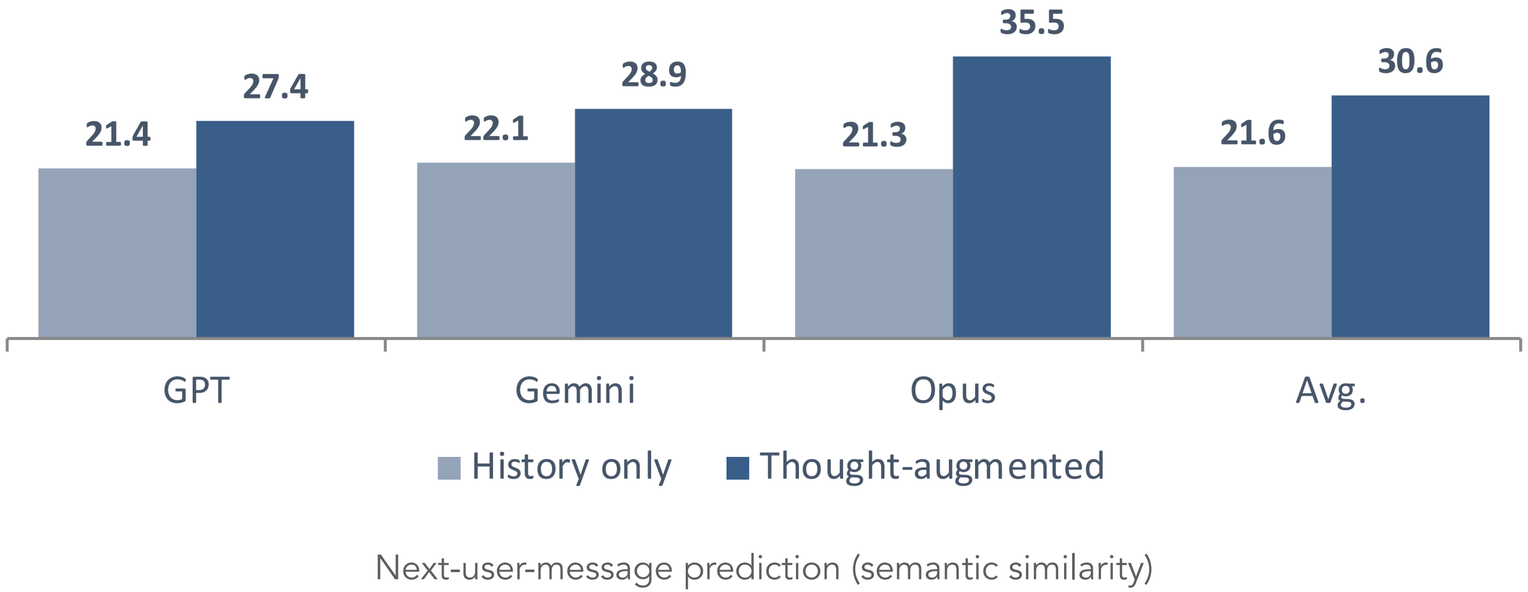
WHY IT MATTERS · 01 · THE RESULT

Thoughts predict **what the user does next**.

+41.7%

relative gain in
next-message prediction

21.6 → 30.6



Next-user-message prediction (semantic similarity)

Given the user's thoughts at inference time, GPT, Gemini, and Opus all predict the next message far more accurately.

Model alignment: a thought becomes a **preference pair** — User-Guided Rewrites, guided by thoughts instead of messages.

PROMPT

User I'm flying to Brazil for a conference in April. What should I prepare?

Assistant A practical travel checklist: passport, visas, flights, hotel…

User Now add what I need for the conference itself.

Prompt = (User 1, Assistant 1, User 2)

WHY IT MATTERS · 02 · THE SETUP

A thought becomes a **preference pair** — User-Guided Rewrites, guided by thoughts instead of messages.

PROMPT

User I'm flying to Brazil for a conference in April. What should I prepare?

Assistant A practical travel checklist: passport, visas, flights, hotel…

User Now add what I need for the conference itself.

Prompt = (User 1, Assistant 1, User 2)

REJECTED · ASSISTANT TURN 2

Conference items: business cards, laptop and chargers, notebook, formal outfit, adapter, snacks, water bottle…

THE USER'S THOUGHT

"Too generic and cluttered — and it misses that this is for a conference."

WHY IT MATTERS · 02 · THE SETUP

A thought becomes a **preference pair** — User-Guided Rewrites, guided by thoughts instead of messages.

PROMPT

User I'm flying to Brazil for a conference in April. What should I prepare?

Assistant A practical travel checklist: passport, visas, flights, hotel…

User Now add what I need for the conference itself.

Prompt = (User 1, Assistant 1, User 2)

REJECTED · ASSISTANT TURN 2

Conference items: business cards, laptop and chargers, notebook, formal outfit, adapter, snacks, water bottle…

THE USER'S THOUGHT

"Too generic and cluttered — and it misses that this is for a conference."

↓

CHOSEN · THOUGHT-GUIDED REWRITE

For the conference: printed slides plus a USB backup, business cards, one formal outfit for your talk — the rest is already above.

Exactly the **User-Guided Rewrites** recipe from Part I — except the guidance is the thought the user never sent.

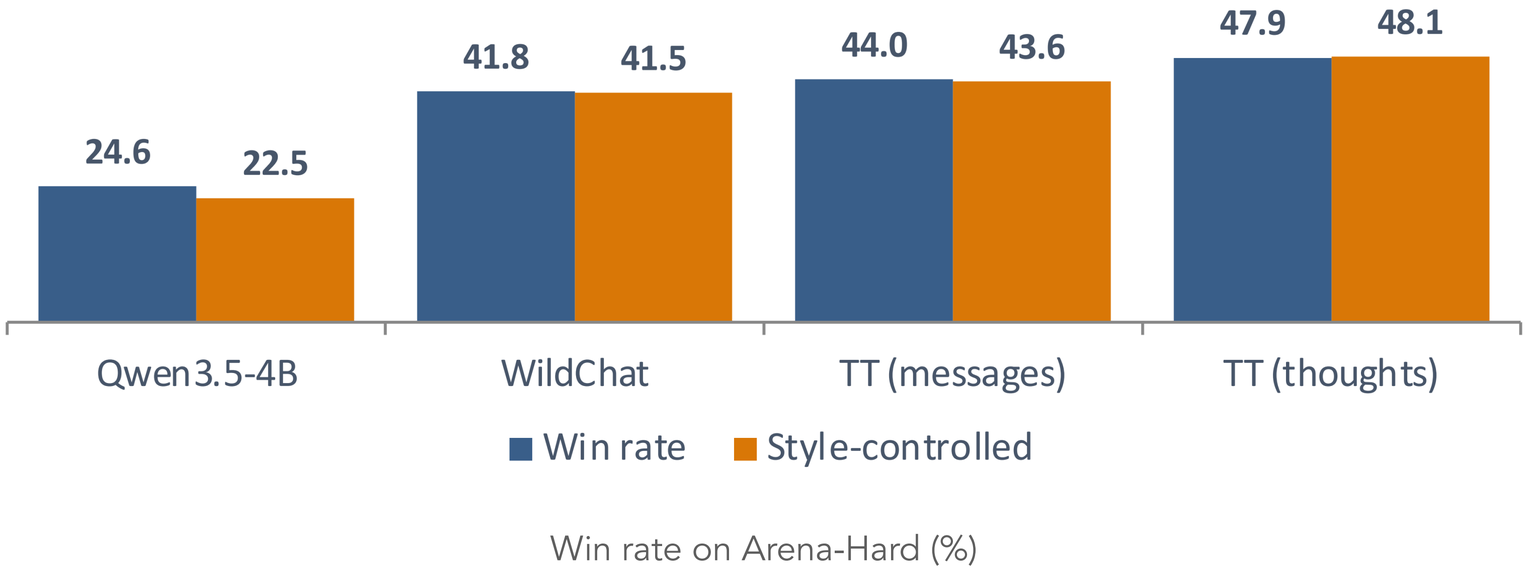
WHY IT MATTERS · 02 · THE RESULT

Thoughts are a better **training signal** for alignment.

+25.6%

absolute gain in
style-controlled win rate

22.5 → 48.1



Win rate on Arena-Hard (%)

Thought-guided rewrites beat message-guided rewrites by **4.5 points** and the WildChat baseline by **6.6** — thoughts carry dissatisfaction users never state out loud.

TAKEAWAYS

ThoughtTrace in three lines.

01

Real conversations paired with users' own thoughts.

first at scale · reasons & reactions · 1,058 users · 20 models

02

Distinct, hard to infer, structured, stage-dependent.

four properties of a genuinely complementary modality

03

And they're actionable — thoughts move real metrics.

+41.7% behavior prediction · +25.6% model alignment

Toward models that keep learning from the people they serve.

User modeling

What users think, and how context shapes it — including personas that update continually.

Model training

Thoughts and organic feedback as supervisory signal, online rather than from static logs.

Evaluation

Thought-centered measures of user satisfaction, plus privacy-preserving personalization.

Thoughts are one modality. The larger question is how a model keeps improving from everything a user does.

NEXT STEPS

Continual learning from **human interaction**.

Learn online

Update from every conversation as it happens — not from frozen logs.

Personalize over months

Personas that persist, update, and transfer across tasks.

Beyond individual chat

From single users to teams, and from chat to autoresearch — learning from human researchers.

In ten years, everyone could have a personal superintelligence that learns from their interactions and empowers what they value.

Thank you

Learning from real human interaction — what users do, and what they think.

RLHI

arxiv.org/abs/2509.25137

ThoughtTrace

arxiv.org/abs/2605.20087