
Learning User Simulators with Turing Rewards

Yingshan Susan Wang^{1*}, Cedegao E. Zhang^{1*}, Linlu Qiu^{1*}, Zexue He^{2*},
 Pengyuan Li³, Alex Pentland^{1,2}, Roger P. Levy¹, Yoon Kim¹

¹Massachusetts Institute of Technology, ²Stanford University, ³MIT-IBM Watson AI Lab
 {susanw26, cedzhang, linluqiu}@mit.edu, zexueh@stanford.edu

Abstract

Learning to simulate human users in interactive settings could advance the training of agent assistants, evaluation of personalization systems, research in the social sciences, and more. Existing approaches generally do so by training a large language model (LLM) to match a single ground truth response, either by maximizing the log probability or by using a similarity reward. We instead propose Turing-RL: a Turing-Test-based reinforcement learning approach for training user simulator models. Turing-RL uses a discriminative Turing reward with an LLM judge to score how indistinguishable a generated response is from the real user’s given the user’s history, and the user simulator LLM learns to produce responses indistinguishable from what the user could have said with such rewards. Across two different domains—conversational chat and Reddit forum discussion—we find that Turing-RL consistently outperforms baseline methods on both LLM and human evaluation metrics. Our study suggests that optimizing for indistinguishability, rather than response matching, is effective for learning user simulators.¹

1 Introduction

The currently dominant use cases of large language models (LLMs) involves having them act as helpful assistants to human users. However, a growing range of applications would benefit from their playing the opposite role—i.e., not the assistant, but the user. Such user simulators could serve as foundational building blocks for social world models in AI agents (Rabinowitz et al., 2018; Collins et al., 2024), training environments and testbeds for interactive systems (Abdulhai et al., 2025; Mehri et al., 2025), and proxies for studying human behaviors at scale (Aher et al., 2023; Park et al., 2023; Lu et al., 2025).

Simulating an individual is however a fundamentally difficult task. What distinguishes one person from another resists easy categorization: two people with identical demographics can hold sharply different opinions (Hwang et al., 2023; Santurkar et al., 2023), and individual preferences cannot be recovered from group-level labels alone (Kirk et al., 2024; Jiang et al., 2025). Recent work has begun addressing this challenge through purpose-built user language models (Naous et al., 2025), latent user-state alignment (Wu et al., 2026), and log-probability maximization with chain of thought (Gandhi et al., 2026). These approaches share a common assumption: the training signal is derived from matching a specific ground truth response, whether by scoring similarity against it with an LLM judge or by maximizing its log-probability. However, the set of plausible responses to a given context is enormous, i.e., the same person in the same context could say many different things. An ideal user simulator model should therefore go beyond replicating the ground truth response and instead produce responses indistinguishable from what the user could have said.

The classic Turing Test (Turing, 1950) operationalizes human-likeness through indistinguishability: a machine passes if an evaluator cannot tell its responses from a human’s.

*Core contributors.

¹Code is available at: <https://github.com/SusanWYS/turing-rl.git>

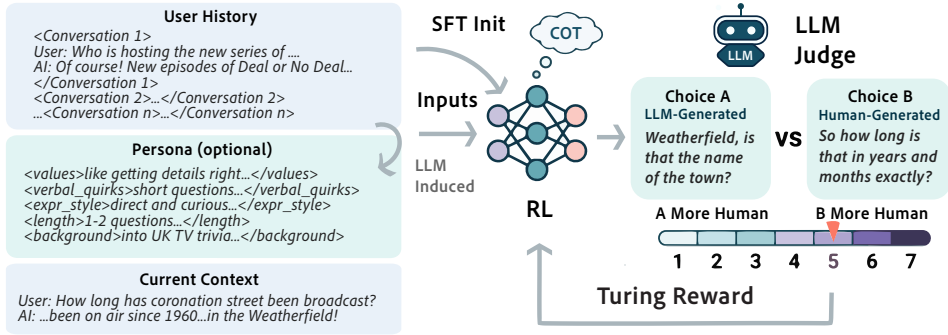


Figure 1: Overview of Turing-RL. Given a user’s history, induced persona, and current conversation context, an SFT-initialized policy generates multiple candidate responses via chain-of-thought (CoT) reasoning. A Turing judge (LLM) compares each candidate against the ground truth human response on a 1–7 scale, scoring which is more likely written by the real user. This discriminative Turing reward is used to train the policy with GRPO.

This is precisely the criterion we want from a user simulator, and it translates directly into a training signal. In this paper, we therefore propose to train LLM user simulators via reinforcement learning (RL) on a *discriminative Turing reward*, wherein an LLM judge assigns a high score when a generated response is deemed human-like, conditioned on the user’s history, which includes the current session history and the user’s behaviors in previous sessions. Figure 1 shows an overview of our approach.

We evaluate our Turing-RL-trained LLMs across two settings that differ substantially in structure, in particular multi-turn dialogue (Kirk et al., 2024) and Reddit forum discussions (Chang et al., 2020). We compare against baseline training signals that have been previously explored in the literature: a response-similarity reward adapted from HumanLM (Wu et al., 2026), and log-probability maximization with chain-of-thought adapted from Gandhi et al. (2026). We find that Turing-RL consistently outperforms both alternatives across both LLM and human evaluation metrics, suggesting that optimizing for indistinguishability is an effective path towards learning user simulators.

2 Learning User Simulators

2.1 Problem Formulation

We study user simulation in an interactive setting, where we are given the current session context x (i.e., prior interactions in the current conversation session or thread) and a user representation u , and need to generate a response y that could have been plausibly produced by the user.

The most straightforward way to represent a user is through their behavior history h (e.g., prior utterances or posts that are not part of the current session). For each user, we reserve a fixed block of k prior interactions at dataset construction time and reuse it as h across all prediction targets for that user. We require h to be disjoint from the current target context x and ground truth response y_* , ensuring that the user representation is never contaminated by the target example. The size k is sampled per user from a small range.

Additionally, we can also pair the raw behavior history with an induced persona ρ . We induce ρ by prompting an auxiliary language model to summarize stable traits from the same user’s history block h . See Figure 7 of the appendix for the prompt used for persona induction. We use $u = (h, \rho)$ as the default in most of our experiments, and ablate on the choice of u (i.e., without the persona) in the ablation study.

2.2 Turing-RL

We use a discriminative Turing reward as the training signal. An LLM judge is shown the user’s history (x, h) alongside two responses to the same context: one written by the real user from the ground truth distribution $y_* \sim p_*(\cdot|x, h)$ and one sampled from the model $y \sim p_\theta(\cdot|x, u)$. The judge rates on a 1–7 Likert scale which response was written by the human, considering the local context, the user’s motive, and stylistic fits. A score of 1 means that the judge deems the human-generated response y_* to be more likely to be written by the human, while a score of 7 means the judge deems the model-generated response y is more likely to be written by the human.² Rather than optimizing for content overlap with a specific ground truth response, this signal trains the model toward the broader quality of being indistinguishable from the target user. The full judge prompt is given in Appendix E.

Letting $s(y, y_*) \in \{1, \dots, 7\}$ be the score assigned by the LLM judge, we cap and normalize the score to $[0, 1]$ to obtain the Turing reward

$$r_{\text{turing}}(y, y_*) = \frac{\min\{s(y, y_*), 5\} - 1}{6}.$$

The reward cap at 5 mitigates against reward hacking: since we would like to produce responses that are human-like, a model that produces responses that are substantially “more human” (as deemed by the judge) than actual human responses would be undesirable and be potentially indicative of reward hacking.³ The final RL objective is then given by

$$\max_{\theta} \mathbb{E}_{y \sim p_\theta(\cdot|x, u), y_* \sim p_*(\cdot|x, h)} [r_{\text{turing}}(y, y_*)],$$

and we train with Group Relative Policy Optimization (GRPO; Shao et al., 2024), after an initial supervised finetuning (SFT) phase on a disjoint subset of the training data.

3 Experimental Setup

3.1 Datasets

We evaluate our approach on two domains: multi-turn chat and Reddit forum discussion. Both domains contain interaction data annotated with user information. We reserve a fixed block of prior interactions as behavior history h , and use the remaining interactions as prediction targets. We split users into training and evaluation sets. Full preprocessing details and statistics are given in Appendix A.

(Chat) Multi-turn chat: PRISM. The PRISM Alignment Dataset (Kirk et al., 2024) contains multi-turn conversations between humans and LLM assistants, spanning 1,500 participants from 75 countries. We select users with at least 6 conversations (1,288 users) and hold out 128 users (880 target user-response turns) for evaluation.

(Reddit) Online forum discussion: ConvoKit. The ConvoKit subreddit corpus (Chang et al., 2020) contains discussions on Reddit. We select 14 subreddits, selecting users with at least 8 threads (1,282 users), and hold out *r/tifu* and *r/worldnews* for evaluation (102 users, 267 examples), with no user overlap with training. Each target thread yields one example: the user’s last comment, and the ancestor comment chain starting from the original post to the target comment.

3.2 Training

SFT warm start. Before RL, we warm-start the policy with SFT on ground truth user responses augmented with chain-of-thought reasoning traces, a common practice in LLM

²In practice, we randomize the ordering of the responses given to the judge to mitigate against ordering bias. We then convert the randomized ordering to the standardized ones as described above.

³We indeed found this type of reward hacking to occur in practice in preliminary experiments.

RL (e.g., Guo et al., 2025). For each training example, we use the Qwen3-8B instruct model to generate a reasoning trace given the context and ground truth response, explaining what could have led the user to write that response. The SFT target pairs the trace with the ground truth: `<reasoning>t</reasoning> [HUMAN]: y`. The model is trained with LoRA using completion-only loss. At inference time, the model generates both the trace and the response autoregressively. See Appendix C.1 for the trace elicitation prompt.

RL training. We optimize the Turing discriminative judge signal with GRPO (Shao et al., 2024). We sample 4 candidate responses per training example and compute advantages via within-group normalization. All models are initialized from the SFT checkpoint discussed above and trained with LoRA (rank $r=64$, scaling $\alpha=32$) for 3 epochs (Hu et al., 2021; Dettmers et al., 2023). GRPO training uses a dataset disjoint from the SFT training set. We additionally apply *length penalty term* that penalize responses falling outside a tolerance band around the ground truth length (see Appendix C.3). We use Qwen3.5-397B-A17B (Qwen Team, 2026) as the judge. Training details are given in Appendix C.

3.3 Baselines

We use Qwen3-8B with thinking mode disabled as the base model, from which we learn a user simulator model through our pipeline. We compare our method with two RL-based methods using alternative reward signals described below. To ensure fair comparison, all RL-based methods are initialized from the same SFT checkpoint, make use of the same history/personas, and are trained with GRPO. We also report the performance of the SFT-init checkpoint, base Qwen3-8B model, Qwen3.5-397B-A17B, and OpenAI GPT-5 as references.

Sim-RL: response similarity reward. Following Wu et al. (2026), this baseline trains with a reward that measures how well the generated response y captures the content of the ground truth response y_* . An LLM judge is given the user’s history, current context, and both responses, and produces an overall alignment score $r_{\text{sim}}(y, y_*) \in [0, 1]$ that reflects semantic similarity while penalizing unsupported content (see Appendix Figure 17 for prompt). This reward encourages the model to say roughly *what* the user said by matching the key points in ground truths.

Logprob-RL: log-probability reward. Gandhi et al. (2026) propose training dialogue simulators by using the log-probability of the ground truth response under the model as the reward, alongside chain-of-thought reasoning. The model generates a reasoning trace z and a candidate response y , and the reward is given by

$$r_{\text{logprob}}(z) = \frac{1}{|y_*|} \log p_{\theta}(y_* | x, u, z),$$

where $|y_*|$ is the number of tokens in the ground truth response and z is the reasoning trace generated by the model. This reward roughly maximizes a lower bound on the log marginal likelihood of the ground truth response.⁴ In their full objective, this reward is combined with an auxiliary SFT term; in our implementation, we drop the auxiliary SFT loss term.⁵

3.4 Evaluation

3.4.1 LLM-as-a-Judge Evaluation

We use LLM-as-a-judge to evaluate generated responses along three dimensions described below. Claude Sonnet 4.6 (Anthropic, 2026) is used for all evaluations. See Appendix E for the full judge prompts and Appendix B for response generation configuration.

⁴Concretely, without the normalization term $\frac{1}{|y_*|}$, we have

$$\mathbb{E}[r_{\text{logprob}}(z)] \leq \log \sum_z p_{\theta}(y_* | x, u, z) p_{\theta}(z | x, u) = \log p_{\theta}(y_* | x).$$

⁵We found that dropping the SFT loss term was more stable for Qwen3-8B GRPO training.

Turing distinguishability. A judge is presented with the user’s behavioral history, the interaction context, and two responses: the ground truth human response and a model-generated candidate, with position randomized. The judge rates which response was written by the real user on a 1–7 scale across three criteria (immediate target, human goal, and communication style). To control for position bias, we evaluate each pair in both orderings and average the scores. Higher scores indicate the generated response is more human-like. This metric mirrors our training signal but makes use of a different and potentially more powerful⁶ Sonnet 4.6 judge than the Qwen2.5-397B-A17B judge used for training.

Response similarity to ground truth. We also measure similarity between the generated and ground truth responses, following Wu et al. (2026). An LLM judge scores how well a generated response captures the overall content of the ground truth human response. The judge extracts key points from the ground truth, then scores semantic similarity, penalizing extraneous content, wrong perspective, and source copying. This yields a single score in $[0, 1]$ measuring whether the model said roughly what the real user said; we report this score as a percentage in results.

Context and user specificity. We finally evaluate whether a generated response is specifically grounded in the current interaction and compatible with the target user, rather than being a generic but plausible reply, which is a common failure of conversational models (Li et al., 2016; Ni et al., 2026). The judge scores two dimensions: *context specificity* (how tightly the response engages with the exact local interaction) and *user evidence compatibility* (how compatible the response is with the target user’s observed behavior without unnaturally exposing or inventing evidence), each weighted 0.5. This yields a single score in $[0, 1]$. For better calibration, we batch judge the scores of multiple responses in one judge call.

3.4.2 Human Evaluation

In addition to LLM-as-a-judge evaluation, we also recruit over three hundred human participants from Prolific to perform a binary-choice Turing test: given a target user’s history and two candidate responses (one real, one model-generated) in randomized order, annotators select which was written by the real user. We evaluate three models: LLM trained with just supervised finetuning (SFT-Init), RL-training with the similarity reward (Sim-RL), and RL-training with our Turing reward (Turing-RL). We do not test Logprob-RL for human evaluation as LLM-as-a-judge results indicate that this variant underperforms Sim-RL and Turing-RL. We target 100 heldout samples per dataset, collecting 600 binary judgments per condition (6 judgments for each target sample) after filtering for comprehension. We report the model win rate (WR), where $WR > 0.5$ means that the model response is picked by human annotators above chance. Full human evaluation details are given in Appendix D.

4 Results

4.1 LLM-as-a-Judge Evaluation

Turing-RL trains human-like simulators. Figure 2 shows Turing judge scores across both domains. Turing-RL outperforms all other models, including Sim-RL and SFT-Init, on both Reddit and Chat. The gap is large on Chat, where Turing-RL exceeds the next-best model by a wide margin.

Notably, user simulation remains a difficult task overall. Even GPT-5 and Qwen3.5-397B—much more capable models than our user simulator—do not improve much compared to Qwen3-8B base. Qualitative inspection reveals that GPT-5 and Qwen3.5-397B tend to produce verbose, overly hedged responses that read as assistant-like rather than human-like (see Figure 6 and Figure 20). Training with the similarity reward does not significantly improve the Turing score over the SFT-Init checkpoint, suggesting that matching the content of the ground truth response does not translate to human-like indistinguishability.

⁶According to some benchmarks, e.g., <https://artificialanalysis.ai/leaderboards/models>.

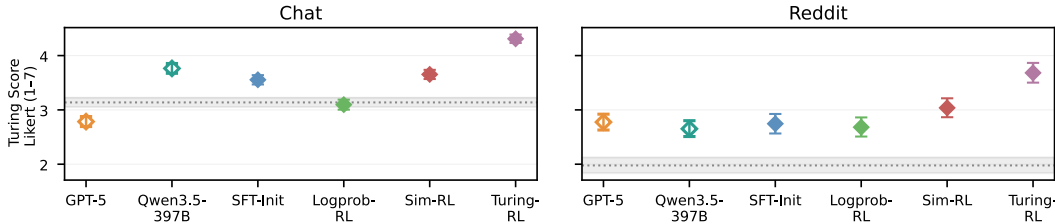


Figure 2: LLMs trained with Turing rewards (Turing-RL) outperforms other training signals on human-likeness in both domains. Turing judge scores (1–7 Likert; higher = harder to distinguish from a real user) from Sonnet 4.6 on Chat and Reddit, with 95% CIs. Hollow markers indicate untrained baselines, while solid markers denote trained Qwen3-8B variants. The dashed line with shaded band marks the Qwen3-8B base model performance with 95% CI.

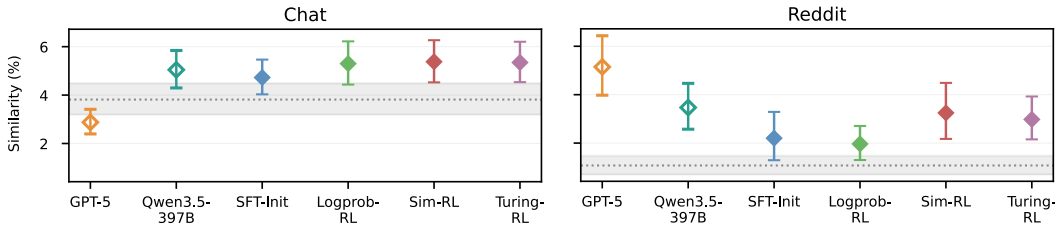


Figure 3: Turing-RL matches Sim-RL even though Sim-RL is explicitly trained to maximize similarity, showing that optimizing for indistinguishability does not sacrifice content alignment. Response similarity to ground truth (Sim, %; higher = more similar to what the user actually said) from Sonnet 4.6 on Chat and Reddit, with 95% CIs. Hollow markers indicate untrained baselines, while solid markers denote trained Qwen3-8B variants. The dashed line with shaded band marks the Qwen3-8B base model performance with 95% CI.

Turing-RL does not sacrifice similarity to ground truth. Figure 3 reports similarity to the ground truth response. Among the trained models, Sim-RL and Turing-RL perform comparably, both improving over the SFT model. This confirms that training with the Turing reward does not sacrifice content alignment: Turing-RL produces responses that are both hard to distinguish from the real user and similar in content to what the user actually said. GPT-5 achieves notably high similarity on Reddit, likely because its verbose outputs happen to cover more of the ground truth key points, even though this verbosity hurts its Turing score.

Turing-RL user simulators are more grounded in context. Beyond matching content, we measure whether generated responses are specifically grounded in the interaction context and compatible with the target user, rather than being generic but plausible replies. Figure 4 shows that Turing-RL and Sim-RL outperform SFT-Init and Logprob-RL in both domains. On Reddit, GPT-5, Qwen3.5-397B, and the ground truth (GT) score highest, while Turing-RL and Sim-RL are close among trained Qwen3-8B models. On Chat, Qwen3.5-397B scores highest overall, followed by Turing-RL, which exceeds both GT and GPT-5. Sim-RL also improves over SFT-Init.

4.2 Human Evaluation

Table 1 reports mean win rates over ground truth responses. These estimates are computed from 100 heldout users per dataset (Reddit/Chat) with 6 selected annotators per heldout user, yielding 600 binary judgments per dataset–model pair. On Chat, Turing-RL has the highest win rate (WR = .57), while SFT-Init and Sim-RL remain close to chance. On Reddit, both Sim-RL and Turing-RL are near chance and much higher than SFT-Init.

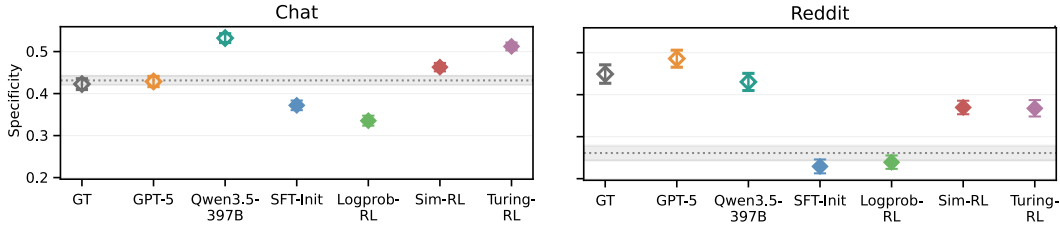


Figure 4: **Turing-RL is among the strongest on Chat, while Turing-RL and Sim-RL are strongest among the trained models on Reddit**, outperforming SFT-Init and Logprob-RL. Response specificity ($[0, 1]$; higher = more grounded in the interaction context and compatible with the target user) from Sonnet 4.6, with 95% CIs. Hollow markers indicate untrained baselines, while solid markers denote trained Qwen3-8B variants. GT is the ground truth human response, and the dashed line marks the Qwen3-8B base model with 95% CI.

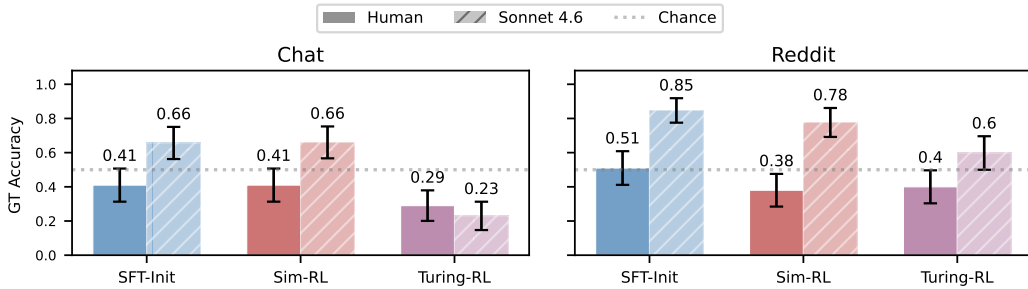


Figure 5: **Comparing human and LLM judge accuracy at identifying the real user’s response.** For each of 50 target users per domain, we take the majority vote from ~ 6 human annotators (solid) and from the Sonnet 4.6 Turing judge. *GT Accuracy* is the fraction of targets correctly identified. Both evaluators agree that Turing-RL is the hardest model to distinguish from real users. **Sonnet 4.6 matches or exceeds human accuracy in most conditions, supporting its use as an automatic evaluation proxy.**

We perform target-level paired permutation tests between Turing-RL and the other two models. On Chat, Turing-RL significantly outperforms both SFT-Init ($p = 0.044$) and Sim-RL ($p = 0.022$). On Reddit, Turing-RL significantly improves on human-likeness compared to SFT-Init ($p = 0.0095$), and no statistically significant difference is detected between Turing-RL and Sim-RL. Further statistical test details are given in Appendix D.5.

We note that Reddit is significantly harder for humans to judge than Chat. As shown in Tables 10 and 11, the Chat/Reddit mean reaction time ratio per target question is 1.43, and the Reddit/Chat mean reaction time per word ratio is 1.48. Therefore, we should treat the human study as a validation of broad trends rather than conclusive evidence, especially for Reddit. Overall, humans agree with LLM judge that Turing-RL-trained user simulators improve on human-likeness compared to SFT-Init on both Reddit and Chat.

Are Humans or LLMs Better Judges? Our human evaluation results are less obviously in favor of the Turing-RL approach. But is it possible that human judges are actually less reliable than the Sonnet 4.6 LLM judge?

	Chat	Reddit
SFT-Init	.49 \pm .061	.41 \pm .045
Sim-RL	.50 \pm .055	.52 \pm .049
Turing-RL	.57 \pm .050	.50 \pm .051

Table 1: Human Turing Test, model win rates against ground truth (\pm 95% CI). Turing-RL greatly improves over SFT-Init on both domains.

User Representation	Chat			Reddit		
	Turing	Sim	Specificity	Turing	Sim	Specificity
$u = (h, \rho)$ (user history + induced persona)	4.31 \pm .08	5.3 \pm .8	.512 \pm .009	3.68 \pm .18	3.0 \pm .9	.367 \pm .020
$u = h$ (user history only)	4.26 \pm .08	4.7 \pm .8	.497 \pm .010	3.78 \pm .18	3.3 \pm 1.0	.364 \pm .019
$u = \rho$ (induced persona only)	4.23 \pm .08	4.1 \pm .7	.509 \pm .009	3.84 \pm .19	2.7 \pm .7	.329 \pm .018

Table 2: Ablation on user representation for Turing-RL. We compare three input conditions: history and persona $u = (h, \rho)$, history only $u = h$, and persona only $u = \rho$. Values are mean \pm 95% CI half-width; Turing is on a 1–7 scale, Sim is reported as a percentage, and Specificity is in $[0, 1]$.

We compare the behavior of the LLM judge and human judges by measuring their accuracy in distinguishing human responses from responses generated by different models. Concretely, we binarize the Sonnet 4.6 Turing score for each evaluation thread (score $< 4 \rightarrow$ GT, $> 4 \rightarrow$ model, $= 4$ excluded) and take the majority vote per target user, mirroring the human forced-choice setup. Figure 5 compares the resulting GT accuracy with human majority-vote accuracy across model on 50 target users for each domain. Human annotators recruited under our experimental setup generally underperform Sonnet 4.6, except for judging Turing-RL on Chat, where the results are similar.

5 Ablations and Analysis

5.1 User Representations

User representation is a core aspect of user simulation. As discussed in §2.1, we condition our models on both the user behavior data h and the induced persona $\rho = f_{\text{LLM}}(h)$ from an auxiliary LLM. How helpful is the induced persona? We ablate on the user representation in both domains and study three different user representations: $u = h$ (history only), $u = \rho$ (persona only), and $u = (h, \rho)$ (both).

Table 2 shows the results. Turing scores are largely robust to the choice of user representation, while specificity is more domain dependent: Chat remains stable across the three inputs, whereas Reddit favors representations that include history. The fact that persona alone achieves comparable Turing scores, but not comparable similarity scores, suggests that the persona captures stylistic and behavioral patterns sufficient for human-likeness, even when it does not help reproduce the exact content of the ground truth response. We also include the normalized Turing reward curves in $[0, 1]$ during GRPO training for the three input conditions, confirming that training dynamics is not much affected by input types (Figure 19).

To mitigate the bias from a single persona inductor model, we further test whether the user representation results depend on the auxiliary model used to induce the persona ρ in Table 12 of the appendix. We find that larger models do not reliably produce better persona for GRPO training, and training with $u = (h, \rho)$ does not necessarily improve performance. We leave more systematic investigation of better user representation as future work.

5.2 Qualitative Examples

Figure 6 shows representative examples from both domains comparing the ground truth, GPT-5, SFT-Init, Sim-RL, and Turing-RL. On Chat, the ground truth is a natural pivot often observed in human-AI conversations. All of the model generations stay anchored to the previous response, and Sim-RL again aligns with GPT-5 in content. Both SFT-Init and Turing-RL ask plausible follow-ups, although the question raised by SFT-Init is already partially answered in the context. On Reddit, the ground truth is a sharp comment. GPT-5 writes an overly verbose response in an assistant-like style. SFT-Init responds by paraphrasing a main point in the context. Sim-RL matches GPT-5 in content but is more succinct. Turing-RL provides a plausible human-like reaction. Additional examples including more baselines are shown in Appendix Figure 20.

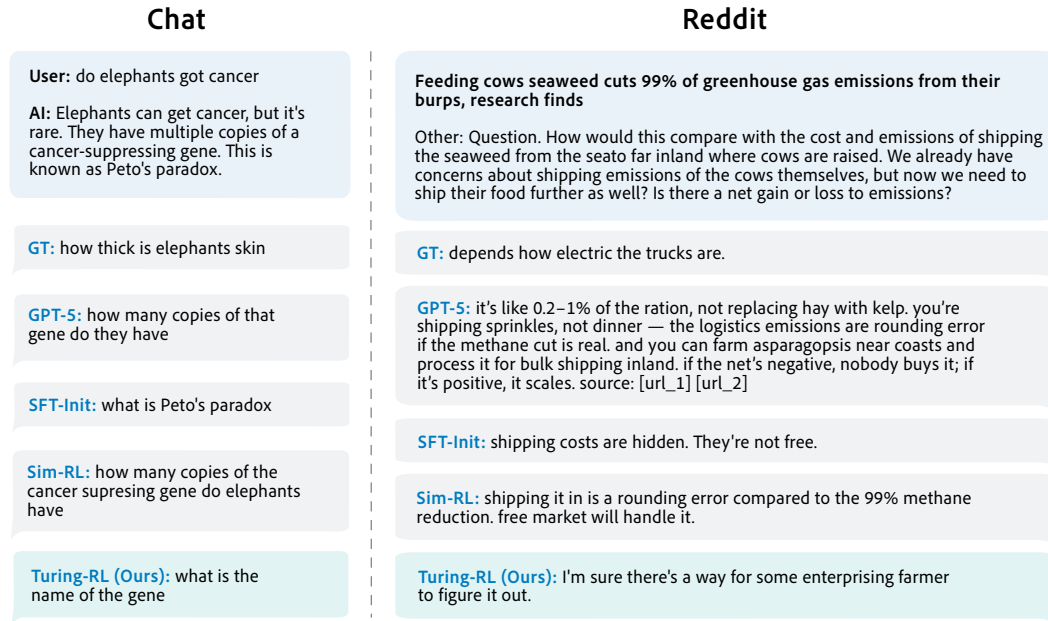


Figure 6: Qualitative comparison of the ground truth, GPT-5, SFT, Sim-RL, and Turing-RL on one Chat and one Reddit example.

6 Discussion

A recurring finding across our experiments is that content matching and human-likeness come apart. While similarity reward improves how well a response covers the ground truth content, it does not necessarily make the response harder to distinguish from the real user. The Turing reward, in contrast, improves indistinguishability without lowering content similarity. This suggests that a discriminative signal is better suited than a matching signal for user simulation.

The same capability that makes a simulator useful, however, also poses risks. A model trained to be indistinguishable from a specific person, conditioned on their prior behavior, is by construction well suited to impersonation; it could be used to fabricate convincing messages attributed to real individuals, or to scale fraud and social-engineering attacks. We emphasize that our simulators are trained and evaluated on public or consented research data and are intended for studying interactive systems and human behavior in aggregate, not for reproducing identifiable individuals. While we believe the positive research and application value of user simulation is substantial and outweigh the risks, it should be pursued alongside safeguards such as watermarking generated content and developing AI generation detectors. As with any dual-use technology, we argue that downstream applications should be evaluated in the context of their specific deployment setting and applicable institutional policies to avoid malicious usage.

There are several directions for future work. First, it would be valuable to study whether user simulators can help LLM assistants become more personalized and better aligned with users' goals and intentions, for example in multi-agent systems (Guo et al., 2024; Tomašev et al., 2026) and cognitive architectures (Sumers et al., 2023; Liu et al., 2026). Second, while this work focuses on training LLMs to produce *outputs* that are indistinguishable from human-written text, it remains an open question whether the models' reasoning *processes* align with those of humans. One avenue is to compare model-generated reasoning traces with verbalized thought traces from humans (Wurgaft et al., 2025; Kargupta et al., 2025); another is to explore richer representations for reasoning beyond natural language alone (Jha et al., 2025; Zhang et al., 2025; Li et al., 2026). Third, reliable user simulators could enable experiments that elicit group and collective behaviors from populations of simulated

agents in open-ended settings, providing a new tool for replicating and extending research in computational social science (Lazer et al., 2009; Ziems et al., 2024).

7 Related Work

LLM-based user simulation. LMs have been used to evaluate dialogue systems (Davidson et al., 2023; Sekulić et al., 2024; Zhou et al., 2026b; Cheng et al., 2025), train conversational agents via self-play (Shah et al., 2018; Abdulhai et al., 2025; Suh et al., 2026), and replicate human subject studies (Aher et al., 2023; Park et al., 2023). Naous et al. (2025) show that assistant-tuned LLMs are structurally misaligned with the user role, while Mehri et al. (2025) address goal drift in task-oriented simulation through explicit state tracking. These works establish that user simulation benefits from purpose-built training but model generic users rather than specific individuals. Beyond standard supervised fine-tuning (Wolf et al., 2019) for user simulation, Abdulhai et al. (2025) use LLM-judged consistency scores as PPO rewards, Wu et al. (2026) optimize for similarity along psychologically grounded dimensions such as stance and emotion, and Gandhi et al. (2026) show that log-probability maximization with latent chain-of-thought outperforms judge-based rewards in generic dialogue prediction. Our discriminative Turing reward takes a different approach: rather than matching content or maximizing likelihood, it trains the model to be indistinguishable from the real user. Concurrent with our work, Zhou et al. (2026a) build foundation models to simulate human behaviors through task-specific and verbal-feedback-based training (Sun et al., 2026), also showing that finetuning a smaller model can beat frontier models on user simulation tasks.

Persona and user representation. It is a common practice to use a natural language description of personality, demographics, and traits to represent a character. Persona-conditioned generation was introduced by Zhang et al. (2018) and extended with pretrained models by Wolf et al. (2019), and Jiang et al. (2024) more recently show that LLMs can reliably express assigned personality profiles in both self-report and free-form writing. Subsequent work enriches user representations: Hwang et al. (2023) show that retrieving relevant past opinions outperforms demographic prompting; Ryan et al. (2025) induce synthetic personas from behavioral traces for personalized reward modeling; and Jiang et al. (2025) demonstrate that individual values cannot be approximated by group-level categories. In this work we combine raw history with an induced persona, providing analysis on different user representations.

Human behavior prediction. Predicting human behaviors, such as people’s choices, actions, and utterances, is a cornerstone task in cognitive and social sciences. Classical approaches model human decision-making include symbolic cognitive architectures (Newell, 1994; Anderson, 2013), the framework of bounded rationality (Simon, 1955), and prospect theory that explains heuristics and biases (Kahneman & Tversky, 2013). Bayesian cognitive science has become a successful paradigm in predicting human behaviors in controlled experimental settings through building structured models that perform probabilistic inference (Griffiths et al., 2024), which has been applied to studying how people simulate the minds of other people (Baker et al., 2009; 2017), the ability known as Theory of Mind (Frith & Frith, 2005). A complementary, data-driven approach trains neural networks on large datasets of human choices, combining cognitive priors and achieving the strong predictive accuracy (Bourgin et al., 2019; Peterson et al., 2021). More recently, LLMs have been evaluated as proxies for human participants in surveys, economic games, and social experiments (Argyle et al., 2023; Horton et al., 2023; Aher et al., 2023; Park et al., 2024). Binz et al. (2025) fine-tune a foundation model on hundreds of psychology experiments to predict trial-level human responses. These approaches have largely focused on constrained, single-turn settings, such as a survey response or an experimental trial.

8 Conclusion

We have proposed training LLM-based user simulators with a discriminative Turing reward, which scores a generated response by how indistinguishable it is from the real user’s response, conditioned on the user’s prior history. Across two substantially different domains, this signal consistently produces more human-like responses than either log-probability maximization or response-similarity rewards under both LLM- and human-based Turing evaluations, without reducing content alignment with the ground truth. Our work suggests a promising recipe for building accurate and generalizable user simulators, with potential applications across a range of downstream settings.

Limitations

We acknowledge the following limitations of our study. First, while we evaluate on two structurally different domains (open-ended chat and forum discussion), there are more extended settings to systematically test the generalization of our Turing reward across other interaction types such as task-oriented dialogue, negotiation, or collaborative problem-solving. Second, our experiments use Qwen3-8B as the base model. While this is sufficient to demonstrate the effectiveness of the Turing reward relative to alternative training signals, it would be valuable to study how our recipe scales to frontier-sized models, where the gap between training signals may narrow or widen in ways we cannot predict from small-model experiments alone. Finally, our discriminative Turing reward relies on a powerful LLM judge (Qwen3.5-397B-A17B) at training time, which introduces both computational cost and a dependence on the judge’s own biases. If the judge has systematic blind spots, the trained simulator may learn to exploit those blind spots rather than achieve genuine human-likeness. Our human evaluation partially addresses this concern by showing that improvements transfer to real human judges, but a more thorough analysis of judge–human disagreement patterns would be valuable for future work.

Acknowledgments

This study was supported by MIT-IBM Watson AI Lab. We thank the Modal credit grant for academics program (<https://modal.com/academics>) for providing some of the cloud computing GPU resources.

References

- Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2511.00222.
- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.
- John R Anderson. *The adaptive character of thought*. Psychology Press, 2013.
- Anthropic. System card: Claude Sonnet 4.6. <https://www.anthropic.com/claude-sonnet-4-6-system-card>, February 2026.
- Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.
- 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.

- 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.
- David D Bourgin, Joshua C Peterson, Daniel Reichman, Stuart J Russell, and Thomas L Griffiths. Cognitive model priors for predicting human decisions. In *International conference on machine learning*, pp. 5133–5141. PMLR, 2019.
- Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 57–60, 2020.
- Myra Cheng, Sunny Yu, and Dan Jurafsky. HumT DumT: Measuring and controlling human-like language in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25983–26008, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1261. URL <https://aclanthology.org/2025.acl-long.1261/>.
- Katherine M. Collins, Iliia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, October 2024. doi: 10.1038/s41562-024-01991-9.
- Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. User simulation with large language models for evaluating task-oriented dialogue. 2023.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314, 2023. URL <https://api.semanticscholar.org/CorpusID:258841328>.
- Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.
- Kanishk Gandhi, Agam Bhatia, and Noah D. Goodman. Learning to simulate human dialogue. *arXiv preprint arXiv:2601.04436*, 2026.
- Thomas L Griffiths, Nick Chater, and Joshua B Tenenbaum. *Bayesian models of cognition: Reverse engineering the mind*. MIT Press, 2024.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.
- John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.
- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL <https://api.semanticscholar.org/CorpusID:235458009>.
- EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*, 2023.
- Kunal Jha, Aydan Yuenan Huang, Eric Ye, Natasha Jaques, and Max Kleiman-Weiner. Modeling others’ minds as code. *arXiv preprint arXiv:2510.01272*, 2025.

- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Person-allm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 3605–3627, 2024.
- Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2410.03868.
- Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.
- Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L Griffiths, Max Kleiman-Weiner, Jiawei Han, et al. Cognitive foundations for reasoning and their manifestation in llms. *arXiv preprint arXiv:2511.16660*, 2025.
- Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.16019.
- David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.
- Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, et al. Simulating society requires simulating thought. *Advances in Neural Information Processing Systems*, 38, 2026.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 110–119, 2016.
- Ryan Liu, Dilip Arumugam, Cedegao E Zhang, Sean Escola, Xaq Pitkow, and Thomas L Griffiths. Cognitive models and ai algorithms provide templates for designing language agents. *arXiv preprint arXiv:2602.22523*, 2026.
- Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Jessie Wang, Qi He, and Dakuo Wang. Can LLM agents simulate multi-turn human behavior? Evidence from real online customer behavior data. *arXiv preprint arXiv:2503.20749*, 2025.
- Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Yisi Sang, Zheshen, Wang, Qi He, and Dakuo Wang. Can llm agents simulate multi-turn human behavior? evidence from real online customer behavior data, 2026. URL <https://arxiv.org/abs/2503.20749>.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.
- Shikib Mehri, Xuhan Yang, Tae 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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2510.06552.
- Allen Newell. *Unified theories of cognition*. Harvard University Press, 1994.

- Bo Ni, Yu Wang, Leyao Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Luera, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen K. Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, Zhengzhong Tu, Alexa Siu, Zichao Wang, Seunghyun Yoon, Nedim Lipka, Namyong Park, Zihao Lin, Trung Bui, Yue Zhao, Tyler Derr, and Ryan A. Rossi. A survey on LLM-based conversational user simulation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4266–4301, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.200. URL <https://aclanthology.org/2026.eacl-long.200/>.
- Joon Sung Park, Joseph C. 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 (UIST)*, 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.
- Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.
- Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.
- Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.
- Michael J Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2506.05598.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.
- Ivan Sekulić, Nicola Ferraro, and Mohammad Aliannejadi. Reliable LLM-based user simulator for task-oriented dialogue systems. 2024.
- Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gökhan Tür. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Industry Papers*, pp. 41–51, 2018.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *Proceedings of the Twentieth European Conference on Computer Systems*, 2024. URL <https://api.semanticscholar.org/CorpusID:272987758>.
- Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pp. 99–118, 1955.
- Joseph Suh, Ayush Raj, Minwoo Kang, and Serina Chang. Quantifying the utility of user simulators for building collaborative llm assistants, 2026. URL <https://arxiv.org/abs/2605.09808>.

- Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.
- Weiwei Sun, Xuhui Zhou, Jiarui Liu, Weihua Du, Haojia Sun, Yiqing Xie, Qianou Ma, Sihao Chen, Mengting Wan, Longqi Yang, Pei Zhou, Sherry Wu, Sean Welleck, Graham Neubig, Yiming Yang, and Maarten Sap. Reinforcing human behavior simulation via verbal feedback, 2026. URL <https://arxiv.org/abs/2605.20506>.
- Nenad Tomašev, Matija Franklin, and Simon Osindero. Intelligent ai delegation. *arXiv preprint arXiv:2602.11865*, 2026.
- A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, March 2020. URL <https://github.com/huggingface/trl>.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. TransferTransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, 2019.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6/>.
- Shirley Wu, Evelyn Choi, Arpandeeep 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.
- Daniel Wurgaft, Ben Prystawski, Kanishk Gandhi, Cedegao E Zhang, Joshua B Tenenbaum, and Noah D Goodman. Scaling up the think-aloud method. *arXiv preprint arXiv:2505.23931*, 2025.
- Cedegao E Zhang, Cédric Colas, Gabriel Poesia, Joshua B Tenenbaum, and Jacob Andreas. Code-enabled language models can outperform reasoning models on diverse tasks. *arXiv preprint arXiv:2510.20909*, 2025.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2204–2213, 2018.
- Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. Odyssim: Building foundation models for human behavior simulation. 2026a.
- Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. Mind the sim2real gap in user simulation for agentic tasks, 2026b. URL <https://arxiv.org/abs/2603.11245>.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.

A Dataset Details

Each training example is a tuple (u, x, y) where the user representation $u = (h, \rho)$ is fixed per user and disjoint from both the current context x and the ground truth response y_* . The behavioral history h is selected once per user from a deterministic seed; the persona ρ is induced from h alone. This separation guarantees that neither h nor ρ leaks information about the target.

A.1 Dataset Preprocessing

We show the dataset statistics after preprocessing in Table 3 below.

	PRISM	ConvoKit
Users	1,288	1,282
History	2–4 conv.	2–6 threads
SFT train rows	3,272	4,264
GRPO train rows	4,174	4,737
GRPO val rows	705	1,661
Test rows	880	267

Table 3: Dataset statistics after preprocessing. History counts are sampled per user. PRISM and ConvoKit use disjoint SFT, GRPO, and held-out test users; the held-out ConvoKit test set comes from `r/tifu` and `r/worldnews`.

(Chat) PRISM. This dataset is distributed under the cc license. We retain PRISM users with at least six conversations, yielding 1,288 qualified users. We randomly hold out 128 qualified users with a fixed seed, yielding 128 held-out test users and 880 test examples. The remaining 1,160 users are split with the same seed into disjoint GRPO and SFT sets using a 60/40 user ratio: 696 users for GRPO and 464 users for SFT. For each user, 2–4 conversations are reserved as h ; the rest serve as targets. Because PRISM dialogues are linear, each user turn in a target conversation yields one example: x is all preceding turns, and y is the user’s utterance at that turn.

(Reddit) ConvoKit. This dataset is distributed under the MIT License. We reconstruct Reddit threads from comment identifiers and parent pointers. We reserve `r/tifu` and `r/worldnews` as held-out test subreddits and remove their users from both SFT and GRPO training. We use the remaining subreddits for SFT and GRPO training. These include `r/AmItheAsshole`, `r/AskMen`, `r/AskWomen`, `r/business`, `r/changemyview`, `r/Economics`, `r/Frugal`, `r/MaliciousCompliance`, `r/news`, `r/relationship_advice`, `r/relationships`, and `r/TrueReddit`.

Users are deterministically split into disjoint SFT and GRPO sets at a 40/60 ratio. For each user, 2–6 threads are reserved as h , and users with insufficient threads are discarded. Each target thread yields one example: y_* is the user’s final comment, and x is the original post plus the ancestor reply chain leading to that comment.

A.2 Persona Induction

Our default user representation pairs the user’s behavioral history with an induced persona. To induce a persona, we prompt GPT-5.4 nano (temperature 0.2, max output length 2,048 tokens) to summarize each user’s history block h into a fixed persona ρ with five fields: values, verbal quirks, expression style, length prior, and background. The same ρ is reused for every example belonging to that user. Figure 7 shows the induction prompt.

System Message (Persona Induction)
<p>You write compressed first-person persona notes as if the target user wrote them about their own habits. Sound like the user, not like an analyst, therapist, teacher, or policy brief. Output only the requested strict JSON object and nothing else.</p>
Persona Induction Prompt (Part 1/2)
<pre> <target_speaker>[HUMAN]</target_speaker> <selected_history> (reserved history block <i>h</i>, with [HUMAN]/[OTHER] speaker labels and bold-delimited target-user responses) </selected_history> <task> Write a compact persona for [HUMAN] in [HUMAN]'s own voice. Rules: - Write each field in first person, as if [HUMAN] is describing their own habits. - Keep it casual and compressed, not polished or formal. - Use only [HUMAN]'s own messages as evidence. - Capture only stable traits that are useful for predicting future messages across contexts. - Prefer concrete wording habits over abstract summaries. - Do not sound like an analyst describing a person from the outside. - Do not use therapist/assistant/policy-brief language. - Avoid abstract labels such as fairness, accountability, autonomy, respect, boundaries, power dynamics, empathy, nuance, skepticism, advocacy, procedural correctness, or similar summary words unless those exact words are frequent in the history. - Do not infer demographics or sensitive attributes. - Do not infer motives, worldview, morals, personality traits, or values unless they are explicitly and repeatedly stated. - Do not include one-off topical stances, transient emotions, or local interaction goals. </pre>

Figure 7: Persona induction prompt (part 1 of 2). The inducer receives the user’s reserved history h formatted with speaker labels and bold-delimited target-user responses, and outputs a structured JSON persona ρ .

B Generation Details

B.1 Sampling Parameters

The sampling parameters used for user simulation are listed in Table 4.

Model	Backend	Max tokens	Temp.	Top- p	Top- k	Pres. pen.	Rep. pen.
Qwen3-8B	vLLM	2048	0.4 / 0.6	1.0	-1	0.5	1.0
GPT-5	API	1024	-	-	-	-	-
Qwen3.5-397B-A17B	vLLM	1024	0.6	1.0	-1	0.5	-

Table 4: Held-out generation sampling parameters. For Qwen3-8B, Reddit generations use temperature 0.4 and Chat generations use temperature 0.6. All Qwen3-8B rows use one generation per target.

Persona Induction Prompt (Part 2/2)
<p>Rules (continued):</p> <ul style="list-style-type: none"> - If evidence is weak, write `unknown`. - Keep each field concise. - Write each field as short fragments or very short sentences, not polished mini-paragraphs. - Do not quote or reproduce complete sentences from the selected history. - Do not quote or reuse exact words or phrases from the selected history, including verdict labels, slang, catchphrases, or short text fragments. - Do not include literal examples from [HUMAN]'s messages, even single-word examples. - Describe lexical habits generically rather than by repeating exact tokens. - Non-text segments such as punctuation styles may be described, but do not quote surrounding text. - Do not include parenthetical example lists, `e.g.` scaffolding, or full sample sentences from [HUMAN]'s messages. <p>Field descriptions:</p> <ul style="list-style-type: none"> - values: only concrete recurring preferences, irritations, or things I keep siding with; avoid abstract moral language unless it is explicit in the history - verbal_quirks: my observable surface-form habits only, such as how I use interjections, punctuation, grammar looseness, discourse markers, formatting quirks, or short repeated lexical patterns. Do not include exact verdict labels or quoted wording from the history. - expression_style: how I usually come across at the sentence level; keep this concrete and plain, not psychological, evaluative, or therapist-like - length_prior: my usual reply length in plain language, mainly sentence count and rough brevity - background: concrete background cues not already captured above, especially repeated personal experiences, self-disclosed roles, responsibilities, constraints, routines, domain familiarity, sources of information, or recurring points of reference that likely shape future responses; do not restate general values, tone, or reasoning style; use `unknown` if nothing strong remains <p>Output only this JSON object shape:</p> <pre>{ "values": "...", "verbal_quirks": "...", "expression_style": "...", "length_prior" : "...", "background": "..." }</pre> <p></task></p>
Model Output
<pre>{ "values": "...", "verbal_quirks": "...", "expression_style": "...", "length_prior" : "...", "background": "..." }</pre>

Figure 8: Persona induction prompt (part 2 of 2). Continuation of rules and field descriptions, followed by the expected model output schema.

B.2 Prompt Formatting

Each example is rendered as a two-message chat sequence: a system message containing the simulation instruction, the task description, and (when applicable) the induced persona ρ (Figure 9); and a user message containing the behavioral history h followed by the current context x . The model generates a chain-of-thought trace z enclosed in `<reasoning>` tags, then the response y prefixed by [HUMAN]:. Figures 10 and 11 show the user message layout for each dataset.

System Message (User Simulation)
<p>You are simulating a human labeled by [HUMAN]. Your job is to predict what [HUMAN] says as if you are that person.</p> <p>[PERSONA] <persona> <values>...</values> <verbal_quirks>...</verbal_quirks> <expression_style>...</expression_style> <length_prior>...</length_prior> <background>...</background> </persona></p> <p>[TASK] Your task is to predict [HUMAN]'s next message, matching [HUMAN]'s writing intentions, style, vocabulary, and tone. Use all provided information about [HUMAN] and the current context to predict the next message. Target your message to [OTHER] or [OTHER - OP], not to people who are described in the context but are not participants in the thread or conversation. Do not answer as an assistant, analyst, or narrator.</p> <p>Before outputting the final answer, briefly reason about what [HUMAN] would say and think step by step. The generated response should naturally follow from your reasoning. Format your output exactly like this: <reasoning>...</reasoning> [HUMAN]: your response Write any reasoning before `[HUMAN]:` enclosed by the reasoning tags, and write only the final response after `[HUMAN]:`. Do not include reasoning after `[HUMAN]:`.</p>

Figure 9: System message shared across all prompt configurations. The [PERSONA] block is included only in persona-conditioned settings; the [TASK] block specifies the prediction objective and output format.

C Training Details

C.1 SFT Training Details

The user splits reserved for SFT training are listed in Table 3. Each SFT training example pairs a ground truth user response with a generated reasoning trace (Lu et al., 2026). To construct each thinking trace, we run Qwen3-8B with thinking mode disabled and provide the user’s reserved history h , induced persona ρ , current context x , and ground truth response y_* . The prompt asks the model to explain what could have led [HUMAN] to write the response, focusing on the local target, intent, stance, style, and plausible length, while explicitly prohibiting copying or closely paraphrasing the ground truth response. We additionally run a ground truth leakage check and regenerate any trace flagged for exposing the wording of y_* until the check passes. The exact thinking-trace generation prompt is shown in Figure 12. The resulting trace is paired with the ground truth as <reasoning> t </reasoning> [HUMAN]: y_* . The loss is applied only to this assistant completion target, while prompt, history, persona, and contexts are masked out. SFT training is implemented with Transformers, TRL SFTTrainer, and PEFT LoRA/QLoRA adapters (Wolf et al., 2020; von Werra et al., 2020; Mangrulkar et al., 2022). Table 5 lists the hyperparameters for the SFT warm-start stage.

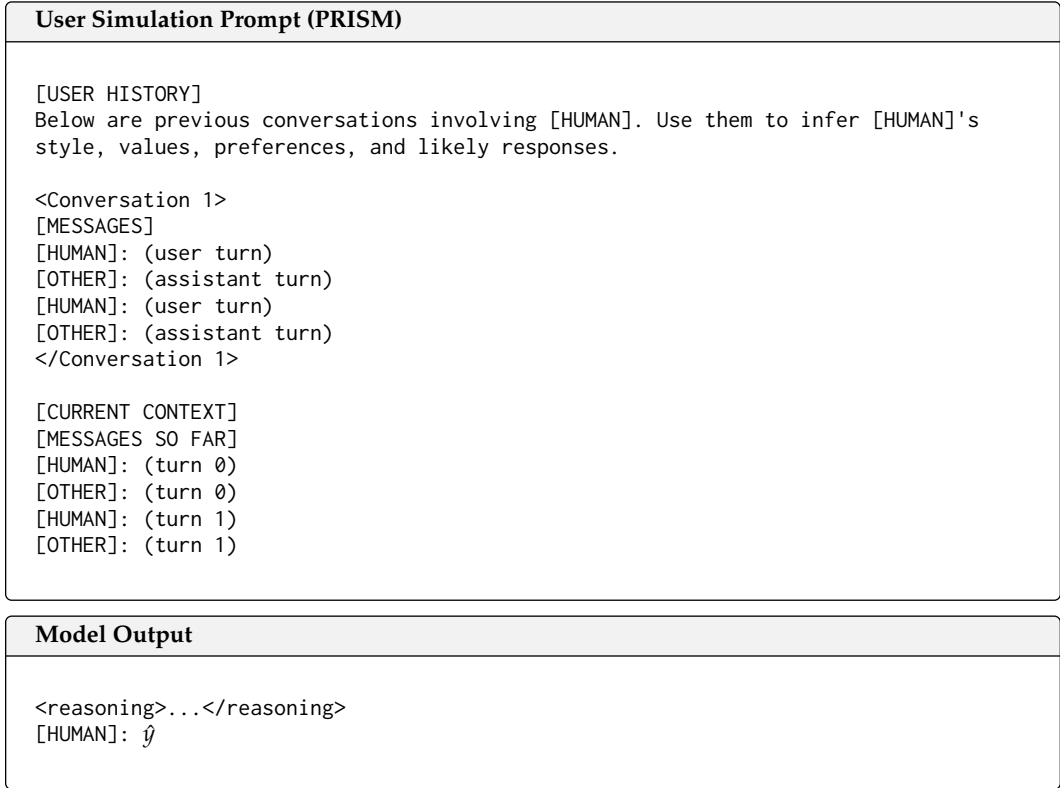


Figure 10: User message layout for PRISM (multi-turn dialogue). History conversations are wrapped in numbered <Conversation> tags; the current context lists turns up to the prediction point.

Parameter	Value
Base model	Qwen3-8B (thinking disabled)
Backend	TRL SFTTrainer + Transformers/PEFT LoRA
LoRA rank / alpha	64 / 128
LoRA dropout	0.05
Learning rate	2×10^{-4}
LR scheduler	cosine
Max sequence length	8192 tokens
Per-device batch size	1
Gradient accumulation steps	16
Number of GPUs	8
Effective batch size	128
Total epochs	3
Precision	bfloat16

Table 5: SFT training hyperparameters.

C.2 GRPO Training Details

Table 6 lists the hyperparameters shared across GRPO variants. All GRPO runs are initialized from the SFT checkpoints and trained on a GRPO user split disjoint from the SFT data (see Table 3). We sample four candidate responses per example, compute group-relative advantages within each candidate group, and optimize the resulting policy objective with LoRA adapters. All GRPO trainings are veRL-based (Sheng et al., 2024). Across SFT and GRPO, we train on B200/B300 machines for about 1680 GPU hours.

User Simulation Prompt (ConvoKit)
<pre> [USER HISTORY] Below are [HUMAN]'s previous messages. <Post 1> [OTHER - OP] (original post text) [OTHER]: (reply) [HUMAN]: (reply) [OTHER]: (reply) </Post 1> [CURRENT CONTEXT] [OTHER - OP] (original post text) [OTHER]: (ancestor reply) [HUMAN]: (earlier comment) [OTHER]: (ancestor reply) [HUMAN]: </pre>
Model Output
<pre> <reasoning>...</reasoning> [HUMAN]: \hat{y} </pre>

Figure 11: Prompt layout for ConvoKit (Reddit forum). History threads are wrapped in numbered <Post> tags; the original poster is marked [OTHER - OP]. The current context shows the OP and ancestor reply chain up to the prediction point.

We optimize the GRPO objective of Shao et al. (2024). For each prompt q , GRPO samples a group of G outputs $\{o_1, \dots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and updates the policy π_{θ} by maximizing

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[\frac{\pi_{\theta}(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{i,t} \right] - \beta \mathcal{D}_{\text{KL}}[\pi_{\theta} \| \pi_{\text{ref}}] \right\} \right],$$

where ε is the clipping range and β the KL coefficient. More specifically:

- $q \sim P(Q)$: a training prompt drawn from the GRPO user split. $P(Q)$ is the empirical distribution over these prompts.
- $\{o_i\}_{i=1}^G$: the $G = 4$ candidate generations sampled per prompt. Each $o_i = (z_i, \hat{y}_i)$ is a reasoning trace z_i followed by the response \hat{y}_i . $o_{i,t}$ is its t -th token.
- π_{θ} : the simulator policy being optimized.
- $\pi_{\theta_{\text{old}}}$: the behavior (sampling) policy that produced the group $\{o_i\}$, namely the policy from immediately before the current update; the per-token ratio $\pi_{\theta} / \pi_{\theta_{\text{old}}}$ is the importance weight. Because we take a single PPO epoch per batch (Table 6), $\pi_{\theta_{\text{old}}} = \pi_{\theta}$ at the first inner step.
- π_{ref} : the SFT checkpoint (Appendix C.1) that initializes RL. The penalty $\mathcal{D}_{\text{KL}}[\pi_{\theta} \| \pi_{\text{ref}}]$ coefficient is $\beta = 1 \times 10^{-3}$ (Table 6).
- ε : the PPO clipping range, set to 0.2 (Table 6).

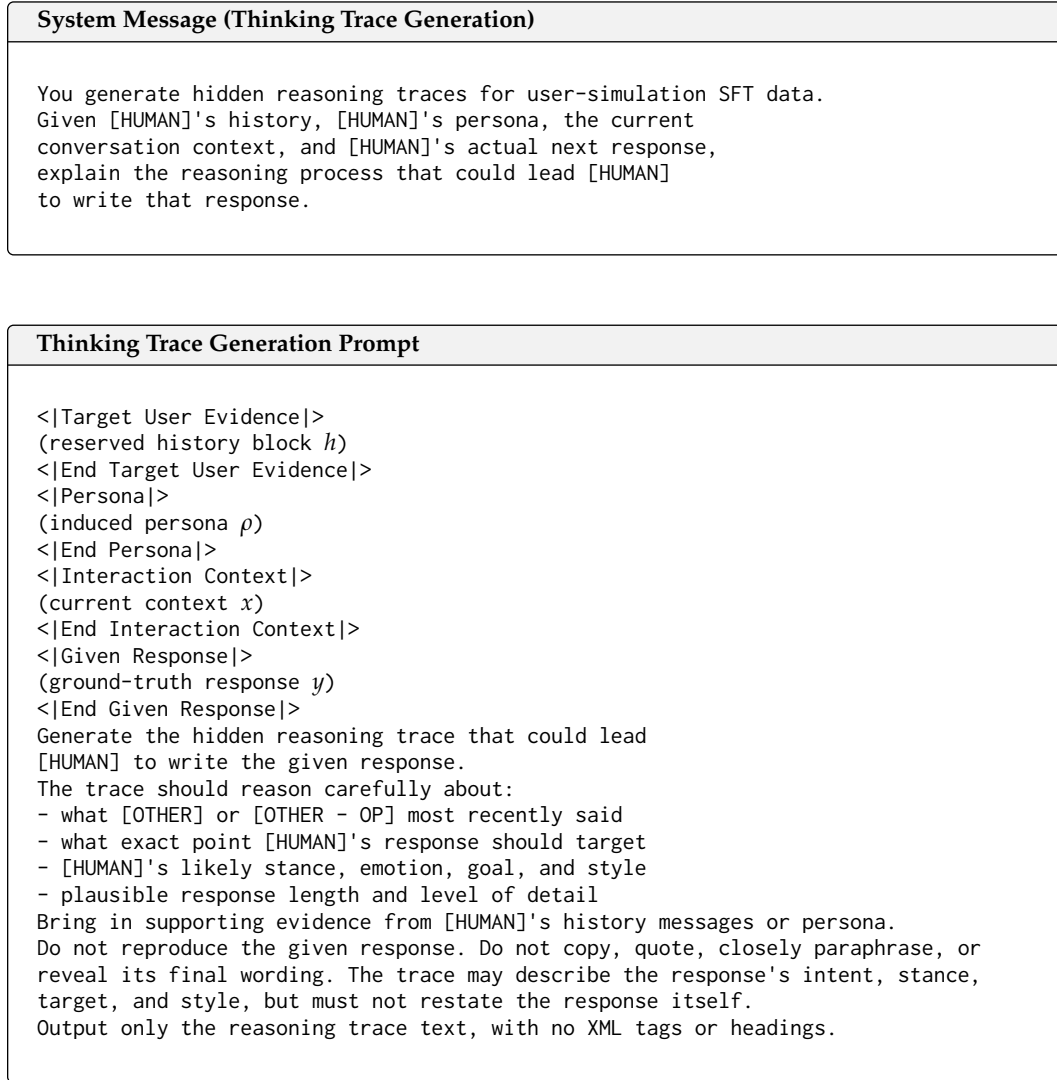


Figure 12: Thinking trace elicitation prompt. A teacher model receives the user’s history h , persona ρ , current context x , and ground-truth response y , and generates a reasoning trace explaining what could lead the user to write that response without reproducing it.

- $\hat{A}_{i,t}$: the advantage of token t in output o_i . $\hat{A}_{i,t} = (r_i - \text{mean}(\mathbf{r})) / \text{std}(\mathbf{r})$, where r_i is individual response reward and $\mathbf{r} = \{r_1, \dots, r_G\}$ are the group rewards. The reasoning tokens z_i receive the same advantage.

C.3 Length Penalty

Our GRPO training incorporates a length penalty for responses that fall outside a tolerance band around the ground truth length. Let ℓ be the generated response length, ℓ_{gt} the ground truth response length, and $r = \ell / \ell_{\text{gt}}$ the length ratio. We define an acceptable length-ratio range $[r_{\text{min}}, r_{\text{max}}]$ within which no penalty is applied. Outside this range, the penalty increases linearly with the relative violation, assigning a larger weight to short responses:

$$p = \min\{\lambda_{\text{short}}v_{\text{short}} + \lambda_{\text{long}}v_{\text{long}}, p_{\text{cap}}\}, \quad (1)$$

Parameter	Value
<i>Shared across all GRPO variants</i>	
Base model	Qwen3-8B (thinking disabled)
Backend	veRL
Temperature	0.6
Top p	1.0
Top k	-1
Initialization	SFT checkpoint
LoRA rank / alpha	64 / 32
Learning rate	1×10^{-5}
KL coefficient (β)	1×10^{-3}
Generations per example (G)	4
Train batch size	128
PPO epochs per batch	1
Total epochs	3
Max response length	1024 tokens
PPO clip ratio	0.2
Loss Aggregation	token mean

Table 6: GRPO training hyperparameters shared across all reward variants.

where $v_{\text{short}} = \max((r_{\text{min}} - r) / r_{\text{min}}, 0)$ and $v_{\text{long}} = \max((r - r_{\text{max}}) / r_{\text{max}}, 0)$. Table 7 lists the dataset-specific parameter values. We set these values based on responses generated by the SFT model.

Parameter	Reddit	Chat
$[r_{\text{min}}, r_{\text{max}}]$	[0.8, 1.1]	[0.6, 1.4]
λ_{short}	0.45	0.40
λ_{long}	0.15	0.20
Penalty cap p_{cap}	0.25	0.40

Table 7: Length penalty parameters for each domain.

D Human Evaluation Details

We conduct a forced-choice Turing test on Prolific. Each participant views a target user’s history and then judges 10 response pairs: one ground truth response and one model-generated response, presented in randomized order. The participant selects which response they believe was written by the real user. Figures 13 and 14 show the annotation interfaces for the Reddit and Chat domains, respectively. Each participant is paid \$6 and the human evaluation costs approximately \$2880 in total (including Prolific service fee and taxes).

D.1 Design

We evaluate three model variants: SFT-Init, Sim-RL, and Turing-RL, on two datasets: ConvoKit (Reddit) and PRISM (Chat). For each dataset, we select 100 target users and split them into 10 groups of 10. Each participant is assigned to one dataset-model-group cell and evaluates the 10 targets in that group. We recruit 6 annotators per cell, yielding 60 participants and 600 binary judgments for each dataset-model pair.

D.2 Comprehension Filtering

Each trial includes comprehension checks that test whether the participant read the target user’s history, such as identifying the user’s most-discussed topic or most recent message. We compute each participant’s comprehension accuracy across all check questions and exclude participants scoring below 75%.

Dataset	Comparison	Mean diff	Paired permutation p
Chat	Turing-RL - SFT-Init	$+.073 \pm .068$	0.044
Chat	Turing-RL - Sim-RL	$+.072 \pm .058$	0.022
Reddit	Turing-RL - SFT-Init	$+.088 \pm .063$	0.0095
Reddit	Turing-RL - Sim-RL	$-.018 \pm .063$	0.61

Table 8: Paired target-level significance tests for Table 1. Mean differences are Turing-RL win rate minus the comparison model’s win rate; \pm values are 95% bootstrap CI.

Dataset	Mean Words/Question	History Mean	Context Mean
Chat	1303.3 ± 140.4	1106.6 ± 129.8	196.7 ± 40.2
Reddit	1676.6 ± 306.4	1328.1 ± 282.5	348.5 ± 63.1

Table 9: Word counts for the displayed human-evaluation stimuli. Values show mean words per target with 95% CI.

D.3 Balancing

After excluding incomplete submissions, defined as submissions with fewer than 10 judgments, and submissions from participants who fail the comprehension filter, we retain the first 6 valid completions for each dataset, model, and user group, ordered by initialization time. This yields exactly 60 participants and 600 votes for each dataset and model.

D.4 Metric

We report the model win rate (WR): the fraction of judgments in which the annotator chose the model-generated response over ground truth. WR is computed per target user (averaging over the ~ 6 annotators for that target) and then averaged across the 100 targets. The 95% confidence interval is $\text{mean} \pm 1.96 \times \text{SEM}$, where SEM is the standard error of the 100 per-target win rates.

D.5 Testing for Statistical Significance

We test pairwise model differences at the heldout-target level. For each dataset and target i , we compute the selected-response win rate for each model and form paired differences $d_i = \text{WR}_{i,\text{Turing-RL}} - \text{WR}_{i,\text{comparison}}$ over the 100 targets per dataset. Table 8 reports the mean paired difference. Bootstrap intervals resample the 100 targets with replacement 10^6 times and recompute the mean difference. The table reports the 2.5–97.5 percentile interval.

We then perform paired permutation test. Under the null that the two models perform equally well, each target-level difference d_i is just as likely to favor either model. For each target with nonzero d_i , we reassign the difference to be $\pm d_i$. Thus, for m nonzero d_i s, there are 2^m possible assignments. For each such assignment, we recompute the mean paired difference. The two-sided p-value is the fraction of assignments whose mean difference has absolute value at least as large as the observed mean difference.

D.6 Word Counts

We summarize the amount of text shown for each target in Table 9, counting words in the displayed user history and target context, excluding response options and interface text.

D.7 Reaction Time

We also record reaction time (RT) on each judgment page, measured from rendering the judgment page to the participant’s click on *Next*. RT includes time spent completing the comprehension check, reading the target context, optionally reopening the user’s history

Metric	Reddit			Chat		
	SFT-Init	Sim-RL	Turing-RL	SFT-Init	Sim-RL	Turing-RL
Mean RT	72.42s \pm 5.57s	73.42s \pm 5.70s	72.63s \pm 6.53s	50.98s \pm 6.42s	56.04s \pm 5.00s	45.82s \pm 3.73s
Median RT	51.36s	53.02s	52.73s	33.87s	38.91s	33.86s

Table 10: Reaction times on the judgment page for the human-evaluation responses. Values show mean RT with 95% CI and median RT. Aggregating across models, the Reddit and Chat mean RTs are 72.83s and 50.95s. The Reddit/Chat mean RT ratio is 1.43.

Metric	Reddit			Chat		
	SFT-Init	Sim-RL	Turing-RL	SFT-Init	Sim-RL	Turing-RL
Mean RT/word	0.188s \pm 0.018	0.183s \pm 0.015	0.171s \pm 0.014	0.120s \pm 0.011	0.133s \pm 0.011	0.130s \pm 0.012
Median RT/word	0.122s	0.134s	0.127s	0.085s	0.101s	0.091s

Table 11: Reaction times normalized by the total history and context word counts for each target user. Values show mean RT per word with 95% CI and median RT per word. Aggregating across models, the Reddit and Chat mean RT/word are 0.185s and 0.125s. The Reddit/Chat mean RT/word ratio is 1.48.

popup, and selecting a response. Table 10 summarizes RT over the selected judgments used in Table 1; Table 11 normalizes the same judgment-page RTs by the displayed words per target.

E Judge Prompts

We use three judge prompts. The Turing distinguishability judge scores how indistinguishable a generated response is from the real user’s response (Figure 16). The similarity judge scores content overlap with the ground truth response (Figure 17). The specificity judge scores whether the response is grounded in the interaction context and compatible with the target user (Figure 18).

F Training Dynamics

GRPO training dynamics of all three rewards along with the input ablation runs for Turing-RL are shown in Figure 19.

G More Qualitative Examples

The qualitative results (ground truth, GPT-5, Qwen3.5-397B, Qwen3-8B Base, SFT-Init, Logprob-RL, Sim-RL, and Turing-RL) for one target each on Reddit and Chat are presented in Figure 20.

User 2's Posting History — Thread 1 of 4
Read through this thread carefully. You can proceed after the timer below.

R/NEWS

ORIGINAL POST (BY OTHER)
Terry Crews Shares His Own Story of Sexual Assault by a Hollywood Executive

OTHER
the floodgates may have just been broken

OTHER
Let's hope. I would love to hear of more powerful executives going down.

OTHER
Don't make the mistake of assuming that this only happens in Hollywood. Remember Trump's comments? The disgusting practice of sexual coercion is undoubtedly rampant wherever huge amounts of power and money reside.

USER 2
"They let you do it."
Makes you wonder how the person who "lets them do it" actually feels.

Back Continue

Conversation Thread User 2's History

User 2's reply appears as "..." below. Which response (A or B) was more likely written by User 2?

R/WORLDNEWS

ORIGINAL POST (BY OTHER)
British ex-spy reveals Kremlin blocked Trump from naming Mitt Romney as secretary of state: report

OTHER
You mean the the guy who wrote the now fully debunked dossier claiming Trump got golden showers in Russia and more bullshit like that great way to vet sources Raw story

OTHER
Swing and a miss!

Actually not a single part of the dossier has been debunked. Nice desperate try though.....

OTHER
actually every part has been debunked
<https://www.forbes.com/sites/paulroderickgregory/2017/01/13/the-trump-dossier-is-false-news-and-heres-why/#466b01666867>
<https://www.washingtontimes.com/news/2018/feb/5/debunked-anti-trump-dossier-deliberately-left-out/>
<http://www.washingtonexaminer.com/fbi-has-not-verified-trump-dossier/article/2641207>
not a single accusation can be verified or has been verified after 18 months , also when your opposition is the one who paid for it the credibility is greatly suspect ... swing and hit

Response A **Response B**

Oh yeah, I'd love to see you apply that to the opposition party who paid for it. Because it's good when we do it and bad when you do it.

Next

Figure 13: User interface used for human annotation in the Reddit domain.

User 1's Conversation History — Conversation 1 of 3

Read through this thread carefully. You can proceed after the timer below.

USER 1
football in england

ASSISTANT
Football in England is extremely popular and has a rich history. The English football league system is one of the best in the world, with teams like Manchester United, Liverpool, and Arsenal being famous worldwide. The Premier League is the top division and is known for its competitiveness and passionate fans.

USER 1
who is the best coach to play in England premier league

ASSISTANT
There have been many great coaches in the English Premier League, and choosing the "best" is subjective. Some notable coaches include Sir Alex Ferguson, Arsène Wenger, José Mourinho, and Pep Guardiola. Each of them has achieved tremendous success and left a significant impact on English football.

USER 1
and what is the best team historically in england

ASSISTANT
Determining the "best" team historically in England is subjective and can vary depending on different criteria such as trophies won, overall success, and cultural impact. Some notable historically successful teams include Manchester United, Liverpool, and Arsenal, each with their own impressive records and legacies in English football.

Back **Continue**

Back **Conversation Thread** **User 1's History**

User 1's reply appears as "..." below. Which response (A or B) was more likely written by User 1?

USER 1
the situation in israel

ASSISTANT
Israel has faced a number of challenges over the past few weeks. The recent wave of protests and violence, which has left over a hundred people injured, has been a source of concern for many. The ongoing situation in the West Bank, including the separation of families, the displacement of communities, and the use of excessive force, has also been a cause for concern. The government has been responding to these challenges by taking various measures, such as the deployment of more security forces, the opening of crossings, and the release of prisoners. Despite these efforts, there is still the potential for further escalation of the situation.

USER 1
...

Before you judge, answer this question about the conversation above:

Which of the following was the last message from the assistant before the target user's reply?

- Here are the top five profitable businesses in the world based on revenue: 1. State Grid Corporation of China 2. Royal...
- Israel has faced a number of challenges over the past few weeks. The recent wave of protests and violence, which has...
- Certainly, here are two potential names for your character, along with explanations for how they can be significant to...

Response A
but why are they deploying more security forces if they are not working

Response B
if there is a war how long do you think it will be

Next

Figure 14: User interface used for human annotation in the Chat domain.

Purpose
 The purpose of this study is to evaluate how well AI language models can imitate the writing style of real online users. Your responses will help us understand the strengths and limitations of current text generation systems.

What You Will Do
 You will review the conversation history of 10 users and, for each user, **judge which of two candidate responses was more likely written by that user**. The study takes approximately **30 minutes**.

Risks
 This study involves reading real conversations between users and AI assistants, which may occasionally contain strong opinions, informal language, or mildly offensive content. There are no other known risks beyond those of everyday computer use. You may stop at any time if you feel uncomfortable.

Benefits
 There are no direct benefits to you. Your participation will contribute to research on language modeling and AI safety.

I agree — begin the study

Task Instructions

In this experiment, you will evaluate **10 different Reddit users**.

For each user, you will:

- 1. Review the user's posting history** — You will see threads where the user has previously commented. The user's messages will be highlighted in **blue**. Take your time reading these to understand how the user writes.
- 2. Read a new conversation** — You will see a conversation where the user's reply is hidden (shown as "...").
- 3. Choose between two responses** — You will be shown Response A and Response B. One of these was actually written by the user. The other was generated by a computer. Your task is to select the response that you believe **was more likely written by the user**, based on their posting history and writing style.

During the judgment, you can click the **"User's History"** button to review the user's previous posts at any time.

You can use the Back button to revisit previous pages within the current user.

I understand, continue

Figure 15: User interface consent and instructions.

Persona Source ($u = (h, \rho)$)	Reddit			Chat		
	Turing	Sim	Specificity	Turing	Sim	Specificity
No persona ($u = h$)	3.78 \pm .18	3.3 \pm 1.0	.364 \pm .019	4.26 \pm .08	4.7 \pm .8	.497 \pm .010
GPT-5.4-nano	3.68 \pm .18	3.0 \pm .9	.367 \pm .020	4.31 \pm .08	5.3 \pm .8	.512 \pm .009
Qwen3-8B	3.47 \pm .19	2.9 \pm 1.1	.337 \pm .020	4.24 \pm .07	4.8 \pm .7	.504 \pm .009
Opus 4.8	3.89 \pm .17	3.2 \pm 1.0	.390 \pm .020	4.19 \pm .08	4.8 \pm .8	.494 \pm .010

Table 12: Ablation on the persona induction model for Turing-RL. We compare the history-only no-persona baseline $u = h$ with history-and-persona models $u = (h, \rho)$ using personas induced by GPT-5.4 nano, Qwen3-8B with thinking enabled, and Opus 4.8. Values are mean \pm 95% CI half-width; Turing is on a 1–7 scale, Sim is reported as a percentage, and Specificity is in $[0, 1]$. **Persona inductor choice has a limited effect relative to confidence intervals**, though Opus performs best on Reddit and GPT-5.4-nano performs best on Chat.

```

Turing Distinguishability Judge (Part 1/10)

## Task

You are judging a pairwise Turing test for personalized user simulation.

Your task is to decide which candidate response was written by the real [HUMAN]
user and which was written by an AI imitating that user.

## Inputs

You will receive:
1. Past responses from [HUMAN]
2. The current conversation context
3. Two candidate responses, Response A and Response B
4. An advisory watchlist for source-copy checks

One candidate is the real [HUMAN] response. The other candidate is AI-generated.

## User History

<|User History|>
{user_history}
<|End User History|>

## Context

<|Context|>
{context}
<|End Context|>

## Candidate Responses

### Response A

<|Response A|>
{response_a}
<|End Response A|>

### Response B

<|Response B|>
{response_b}
<|End Response B|>

## Advisory Watchlist

### Source-Copy Watchlist

<|Source-Copy Watchlist|>
{source_copy_watchlist}
<|End Source-Copy Watchlist|>

## Evaluation Procedure

Evaluate each response independently before comparing them. Reason backward from
the response to the likely target, goal, and style.

Score each response on three criteria from 0.0 to 1.0.

```

Figure 16: Exact Turing distinguishability judge prompt used in the experiments, split across nine parts for typesetting.

Turing Distinguishability Judge (Part 2/10)

Criteria

1. Immediate target

Identify the exact part of the current context that the response addresses.

Consider whether the response:

- Reacts to a specific point in the latest [OTHER] turn, the broader current context, or a plausible topic pivot
- Understands what the other person actually said
- Speaks from [HUMAN]'s perspective rather than [OTHER]'s perspective
- Is directed at [OTHER] or [OTHER - OP], rather than at someone merely described in the thread

A topic pivot can be valid when it is a plausible next move for [HUMAN]. Do not penalize a pivot only because it does not directly answer the previous turn, especially if [HUMAN] often pivots in the history or current context.

Do not reward broad persona fit if the response misstates a speaker's stance, role, or personal experience; assign a low `immediate_target_score`.

`immediate_target_score`:

- 0.0-0.2 = Wrong or absent target. The response does not reply to the current context, targets the wrong person, takes [OTHER]'s role, or responds to someone only described in the thread.
- 0.2-0.4 = Weak target. The response is broadly on topic but does not clearly address the latest [OTHER] turn, a relevant current-context point, or a plausible human topic pivot.
- 0.4-0.6 = Mixed target. The response addresses the general context or makes a possible pivot, but the exact target is ambiguous, unsupported, or only loosely connected to the current exchange.
- 0.6-0.8 = Good target. The response addresses a plausible part of the current context or makes a plausible topic pivot from [HUMAN]'s perspective, with only minor ambiguity.
- 0.8-1.0 = Strong target. The response clearly addresses the most plausible current-context point, or makes a strongly human-plausible pivot, from [HUMAN]'s perspective and directed at [OTHER] or [OTHER - OP].

Turing Distinguishability Judge (Part 3/10)

2. Human goal

Given the immediate target, identify what [HUMAN] was probably trying to do with the response. When judging the goals, use [HUMAN]'s history as evidence, not as a rigid script.

Consider whether the goal:

- Is plausible for [HUMAN] in this context, given [HUMAN]'s history
- Fits the local conversation pattern
- Preserves a specific personal or context-sensitive intent when one is available
- Avoids replacing [HUMAN]'s likely intent with a generic information request, assistant-like task, or narrator behavior

Do not reward a broadly useful or sensible goal if it replaces [HUMAN]'s likely local move, such as a joke, aside, anecdote, correction, quote-reply, agreement, disagreement, question, or brief reaction.

human_goal_score:

- 0.0-0.2 = Wrong or implausible goal. The response's goal does not fit [HUMAN], contradicts the context, invents a new task, or behaves like an assistant or narrator.
- 0.2-0.4 = Weak goal. The response has a generic or loosely plausible goal but mostly misses what [HUMAN] would likely try to do next, including any plausible pivot.
- 0.4-0.6 = Mixed goal. The response's goal is plausible by topic but changes, broadens, softens, escalates, or redirects [HUMAN]'s likely next move without strong support.
- 0.6-0.8 = Good goal. The response's goal, including a reasonable new-question or topic-pivot goal, is plausible for [HUMAN] in this context, with only minor uncertainty or drift.
- 0.8-1.0 = Strong goal. The response's goal naturally follows from [HUMAN]'s current intent, history, or local topic-switching pattern, while preserving personal or context-sensitive intent.

Turing Distinguishability Judge (Part 4/10)

3. Communication style and length

Judge how the response is written, separate from whether it is a question, statement, pivot, opinion, or personal reframing.

Compare the response to [HUMAN]'s past responses and the local framing. Consider:

- Wording and phrasing
- Tone, humor, bluntness, and emotion
- Length and level of detail
- Grammar, capitalization, punctuation, and misspellings
- Specificity and natural level of effort

Do not reward generic fluency. Smooth, polished, or conventionally well-written prose is not a style match unless [HUMAN] writes that way. Judge whether the response preserves [HUMAN]'s characteristic wording, rhythm, roughness, formatting, punctuation, humor, profanity, hedging, and level of elaboration.

Match length as a style feature. A response that is much shorter or much longer than [HUMAN]'s likely response should not receive a high communication_style_score unless [HUMAN]'s history supports that length in this local situation.

Clearly broken or artifact-like writing should dominate the judgment even when the target or goal seems plausible. Hard penalize artifact-like generations.

communication_style_score:

- 0.0-0.2 = Strong mismatch. Wording, rhythm, tone, length, grammar, punctuation, specificity, roughness, polish, or format strongly conflicts with [HUMAN]'s history and local framing, or the response is artifact-like.
- 0.2-0.4 = Weak style match. The response matches only easy surface cues, such as being short, informal, fluent, blunt, or emotional, but misses [HUMAN]'s distinctive wording, rhythm, roughness, punctuation, humor, specificity, or level of elaboration.
- 0.4-0.6 = Mixed style match. Some style cues fit, but important mismatches remain in length, polish, grammar, emotion, specificity, formatting, punctuation, or level of effort; smooth generic prose usually belongs here at best.
- 0.6-0.8 = Good style match. The response mostly matches [HUMAN]'s characteristic wording, rhythm, tone, length, formatting, punctuation, specificity, and level of elaboration, with only minor mismatch; length should be in the same rough range as [HUMAN]'s likely response.
- 0.8-1.0 = Strong style match. The response sounds specifically like [HUMAN], including distinctive phrasing, rhythm, roughness or polish, humor, profanity, hedging, formatting, punctuation, specificity, natural level of effort, and closely matched length.

Turing Distinguishability Judge (Part 5/10)

Penalty Checks

Each penalty is scored from 0.0 to 1.0. Use the same scale for every penalty:

- 0.0 = No issue.
- 0.1-0.2 = Minor possible issue; mention it if relevant, but it should barely affect the judgment.
- 0.3-0.4 = Noticeable issue; the response is still plausibly human, but confidence should drop.
- 0.5-0.6 = Serious issue; the response has a substantial penalty-worthy flaw.
- 0.7-0.8 = Very strong issue; the response is unlikely to be human for this reason.
- 0.9-1.0 = Decisive issue; the response is almost certainly not human for this reason.

Bad quote or source copy

The source-copy watchlist is an advisory 5-gram scan against user history and current context. Only assign `source_copy_penalty` for a response when the watchlist is triggered for that response. If the watchlist is off or not triggered for a response, assign 0.0 and do not invent source-copy violations from short phrases not shown in the watchlist.

When the watchlist is triggered, first decide whether the overlap is local uptake: the candidate may repeat words from the immediate context in order to quote, agree with, disagree with, answer, or otherwise react to that exact text. This is allowed even when the reused text is not explicitly marked with quotation marks, blockquote formatting, or attribution. Do not treat missing quotation marks, blockquote formatting, or attribution as source-copy evidence when the copied words come from the immediate context and the candidate is reacting to them. Do not penalize logical quotes without quotation marks. Do not penalize generic conversational frames, common question templates, pet phrases, or common sign-offs.

If copied text is not a natural quote or local uptake (i.e., a direct copy from the user history), assign a high `source_copy_penalty`. If the response quotes text but the quotation logic is confusing or contradictory, assign a high `source_copy_penalty`.

Turing Distinguishability Judge (Part 6/10)

Wrong target or speaker role

Assign `wrong_target_or_role_penalty` when a response speaks from the wrong role, addresses the wrong speaker, or assigns an unsupported stance, experience, motive, relationship, or conflict to the speaker it addresses.

This includes responses that:

- take another speaker's role instead of [HUMAN]'s role, including by assigning [OTHER]'s or [OTHER - OP]'s views, experiences, motives, relationships, or conflicts to [HUMAN]
- address someone only described in the context instead of an actual current-context speaker, such as [OTHER] or [OTHER - OP]
- address a real current-context speaker but treat that speaker as holding a view they did not state
- assign the addressed speaker a personal experience, motive, relationship, or conflict absent from the exchange

Use a high `wrong_target_or_role_penalty` when the role or target error makes the response implausible as [HUMAN]'s reply.

Unsupported adversarial reframing

Assign `unsupported_adversarial_reframing_penalty` when a response attacks, corrects, rebuts, or cynically reframes a claim that the current-context speakers did not actually make.

Broad persona fit alone is not enough. A response can sound like [HUMAN]'s general argumentative style while still targeting the wrong claim.

Watch for:

- generic reframes such as "the real issue is", "that's not X, that's Y", or "you're not X, you're Y"
- repeated questions, stacked sarcastic questions, or blunt challenges that do not logically follow from the exchange
- unsupported accusations combined with generic demands for evidence, validation, or clarification
- invented motives, conflicts, criticisms, or roles absent from the current context

When a response is framed as a pushback, rebuttal, or challenge, check whether the objection responds to the claim actually made. If it attacks a premise the other speaker did not rely on, misstates a speaker's stance, role, or personal experience, or reframes the issue into a different dispute, assign a high `unsupported_adversarial_reframing_penalty`.

Turing Distinguishability Judge (Part 7/10)**### Assistant-like response**

Assign `assistant_like_penalty` when a response clearly reads like chatbot output rather than an organic human comment based on surface form. Watch for numbered or bulleted action plans, section headers, step-by-step framing, summary or conclusion sections, repeated direct-address coaching, generic reassurance, overgenerated transition phrases such as "But wait" or "But wait-", template phrases such as "You're absolutely right", "You're absolutely correct", "here's what you need to do", "let's break this down", "next steps", or "set boundaries", overuse of dashes "-", semicolons ";", or emojis, especially multiple emojis in a short response when not justified by [HUMAN]'s history or the local context, unnatural colloquial apostrophe abbreviations such as "'em" when not supported by [HUMAN]'s history or the local context, and other AI-written template patterns. Do not treat "[...quoted text...]" as assistant-like by itself; it may appear in human-written responses.

Do not penalize a response merely because it uses bullets, links, citations, dashes, or structured argumentation if that format is natural for the platform, thread type, or [HUMAN]'s history.

Scoring and Rating**Compute:**

```
base_score_a = immediate_target_score_a + human_goal_score_a +
communication_style_score_a
```

```
base_score_b = immediate_target_score_b + human_goal_score_b +
communication_style_score_b
```

```
penalty_a = ((source_copy_penalty_a + wrong_target_or_role_penalty_a +
unsupported_adversarial_reframing_penalty_a + assistant_like_penalty_a) / 4) * 3
```

```
penalty_b = ((source_copy_penalty_b + wrong_target_or_role_penalty_b +
unsupported_adversarial_reframing_penalty_b + assistant_like_penalty_b) / 4) * 3
```

```
response_a_score = max(0.0, base_score_a - penalty_a)
```

```
response_b_score = max(0.0, base_score_b - penalty_b)
```

```
score_gap = response_b_score - response_a_score
```

Convert score_gap to the final 1-7 rating:

- rating = 1 if `score_gap` <= -2.0
- rating = 2 if -2.0 < `score_gap` <= -1.0
- rating = 3 if -1.0 < `score_gap` <= -0.25
- rating = 4 if -0.25 < `score_gap` < 0.25
- rating = 5 if 0.25 <= `score_gap` < 1.0
- rating = 6 if 1.0 <= `score_gap` < 2.0
- rating = 7 if `score_gap` >= 2.0

Rating scale:

- 1 = Definitely A is the real human response
- 2 = Very likely A
- 3 = More likely A than B
- 4 = Cannot tell / equally likely
- 5 = More likely B than A
- 6 = Very likely B
- 7 = Definitely B is the real human response

Turing Distinguishability Judge (Part 8/10)**## Output Format**

Return exactly one valid JSON object with this schema:

```
{
  "immediate_target_a": "<What exact part of the context is Response A reacting to, including any plausible topic pivot? Does it understand the latest turn, preserve the speakers' actual stance/role/personal experience, and avoid answering an invented position or conflict? Is it from [HUMAN]'s perspective and targeted to [ OTHER] or [OTHER - OP]?>",
  "immediate_target_score_a": <number from 0.0 to 1.0>,
  "immediate_target_b": "<What exact part of the context is Response B reacting to, including any plausible topic pivot? Does it understand the latest turn, preserve the speakers' actual stance/role/personal experience, and avoid answering an invented position or conflict? Is it from [HUMAN]'s perspective and targeted to [ OTHER] or [OTHER - OP]?>",
  "immediate_target_score_b": <number from 0.0 to 1.0>,
  "human_goal_a": "<Given the immediate target, what was [HUMAN] probably trying to do with Response A? Is that goal plausible for [HUMAN] in this context, specific or context-sensitive rather than only generic, and does it preserve [HUMAN]'s likely local move such as a joke, aside, anecdote, correction, quote-reply, agreement, disagreement, question, or brief reaction? Do not over-penalize a plausible pivot, blunt statement, opinion, or personal reframing.>",
  "human_goal_score_a": <number from 0.0 to 1.0>,
  "human_goal_b": "<Given the immediate target, what was [HUMAN] probably trying to do with Response B? Is that goal plausible for [HUMAN] in this context, specific or context-sensitive rather than only generic, and does it preserve [HUMAN]'s likely local move such as a joke, aside, anecdote, correction, quote-reply, agreement, disagreement, question, or brief reaction? Do not over-penalize a plausible pivot, blunt statement, opinion, or personal reframing.>",
  "human_goal_score_b": <number from 0.0 to 1.0>,
  "communication_style_a": "<Does Response A's wording, rhythm, tone, length, grammar, roughness or polish, humor, profanity, hedging, formatting, capitalization, punctuation, misspellings, specificity, and level of elaboration match [HUMAN]'s past responses and local framing? Do not reward generic fluency unless [HUMAN] writes that way; note whether length is in [HUMAN]'s likely range.>",
  "communication_style_score_a": <number from 0.0 to 1.0>,
  "communication_style_b": "<Does Response B's wording, rhythm, tone, length, grammar, roughness or polish, humor, profanity, hedging, formatting, capitalization, punctuation, misspellings, specificity, and level of elaboration match [HUMAN]'s past responses and local framing? Do not reward generic fluency unless [HUMAN] writes that way; note whether length is in [HUMAN]'s likely range.>",
  "communication_style_score_b": <number from 0.0 to 1.0>,
  "base_score_a": <immediate_target_score_a + human_goal_score_a + communication_style_score_a>,
  "base_score_b": <immediate_target_score_b + human_goal_score_b + communication_style_score_b>,
  "response_a_score": <number from 0.0 to 3.0>,
  "response_b_score": <number from 0.0 to 3.0>,
  "score_gap": <response_b_score - response_a_score>,
}
```

Turing Distinguishability Judge (Part 9/10)

```

"response_a_source_copy": "<If source-copy watchlist is on for Response A,
explain whether the matched text is a generic frame, common template, pet phrase,
sign-off, natural quote, local uptake, direct copy from user history, or confusing/
contradictory quote. If the watchlist is off, write an empty string.>",
"source_copy_penalty_a": <number from 0.0 to 1.0>,
"response_b_source_copy": "<If source-copy watchlist is on for Response B,
explain whether the matched text is a generic frame, common template, pet phrase,
sign-off, natural quote, local uptake, direct copy from user history, or confusing/
contradictory quote. If the watchlist is off, write an empty string.>",
"source_copy_penalty_b": <number from 0.0 to 1.0>,
"response_a_wrong_target_or_role": "<Explain any wrong speaker role, wrong
addressee, unsupported stance attribution, or unsupported personal experience/
motive/relationship/conflict for Response A. If there is no issue, write an empty
string.>",
"wrong_target_or_role_penalty_a": <number from 0.0 to 1.0>,
"response_b_wrong_target_or_role": "<Explain any wrong speaker role, wrong
addressee, unsupported stance attribution, or unsupported personal experience/
motive/relationship/conflict for Response B. If there is no issue, write an empty
string.>",
"wrong_target_or_role_penalty_b": <number from 0.0 to 1.0>,
"response_a_unsupported_adversarial_reframing": "<Explain any unsupported attack,
correction, cynical reframe, wrong-claim rebuttal, invented motive/conflict/
criticism, or illogical challenge for Response A. If there is no issue, write an
empty string.>",
"unsupported_adversarial_reframing_penalty_a": <number from 0.0 to 1.0>,
"response_b_unsupported_adversarial_reframing": "<Explain any unsupported attack,
correction, cynical reframe, wrong-claim rebuttal, invented motive/conflict/
criticism, or illogical challenge for Response B. If there is no issue, write an
empty string.>",
"unsupported_adversarial_reframing_penalty_b": <number from 0.0 to 1.0>,
"response_a_assistant_like": "<Explain any assistant-like surface-form issue for
Response A: generic chatbot formatting, template phrasing, repeated direct-address
coaching, generic reassurance, overgenerated transition phrases, overuse of dashes,
semicolons, emojis, unsupported colloquial apostrophe abbreviations, or artifact-
like surface form. If there is no issue, write an empty string.>",
"assistant_like_penalty_a": <number from 0.0 to 1.0>,
"response_b_assistant_like": "<Explain any assistant-like surface-form issue for
Response B: generic chatbot formatting, template phrasing, repeated direct-address
coaching, generic reassurance, overgenerated transition phrases, overuse of dashes,
semicolons, emojis, unsupported colloquial apostrophe abbreviations, or artifact-
like surface form. If there is no issue, write an empty string.>",
"assistant_like_penalty_b": <number from 0.0 to 1.0>,
"penalty_a": <((source_copy_penalty_a + wrong_target_or_role_penalty_a +
unsupported_adversarial_reframing_penalty_a + assistant_like_penalty_a) / 4) * 3>,
"penalty_b": <((source_copy_penalty_b + wrong_target_or_role_penalty_b +
unsupported_adversarial_reframing_penalty_b + assistant_like_penalty_b) / 4) * 3>,
"reasoning": "<Concise explanation of the base scores, penalties, final score gap
, topic-pivot evidence, goal specificity, style evidence, capitalization and
punctuation evidence, and final rating.>",
"rating": <integer from 1 to 7>
}

```

Turing Distinguishability Judge (Part 10/10)

Format Rules

- Output only valid JSON.
- Keep every explanation clear and concise.
- All criterion scores must be numbers from 0.0 to 1.0.
- All penalty scores must be numbers from 0.0 to 1.0.
- "base_score_a", "base_score_b", "penalty_a", and "penalty_b" must match the formulas above.
- "response_a_score" and "response_b_score" must be numbers from 0.0 to 3.0.
- "score_gap" must equal response_b_score - response_a_score.
- "rating" must be a single integer from 1 to 7.

Your output:

Model Output

```
{
  "immediate_target_a": "...",
  "immediate_target_score_a": 0.0,
  "immediate_target_b": "...",
  "immediate_target_score_b": 0.0,
  "human_goal_a": "...",
  "human_goal_score_a": 0.0,
  "human_goal_b": "...",
  "human_goal_score_b": 0.0,
  "communication_style_a": "...",
  "communication_style_score_a": 0.0,
  "communication_style_b": "...",
  "communication_style_score_b": 0.0,
  "base_score_a": 0.0,
  "base_score_b": 0.0,
  "response_a_score": 0.0,
  "response_b_score": 0.0,
  "score_gap": 0.0,
  "response_a_source_copy": "",
  "source_copy_penalty_a": 0.0,
  "response_b_source_copy": "",
  "source_copy_penalty_b": 0.0,
  "response_a_wrong_target_or_role": "",
  "wrong_target_or_role_penalty_a": 0.0,
  "response_b_wrong_target_or_role": "",
  "wrong_target_or_role_penalty_b": 0.0,
  "response_a_unsupported_adversarial_reframing": "",
  "unsupported_adversarial_reframing_penalty_a": 0.0,
  "response_b_unsupported_adversarial_reframing": "",
  "unsupported_adversarial_reframing_penalty_b": 0.0,
  "response_a_assistant_like": "",
  "assistant_like_penalty_a": 0.0,
  "response_b_assistant_like": "",
  "assistant_like_penalty_b": 0.0,
  "penalty_a": 0.0,
  "penalty_b": 0.0,
  "reasoning": "...",
  "rating": 4
}
```

Response Similarity Judge (Part 1/3)
<p>You are a helpful and meticulous evaluator. Your task is to score how well the generated response(s) align with the ground truth user response. Description of response: [HUMAN]'s actual written comment or reply text.</p> <p>You will be given past messages for [HUMAN], the current context, the ground truth response, and generated response(s) that you should evaluate.</p> <p>Provided Information: {context}</p> <p>< The Start of Ground Truth Response > {ground_truth} < The End of Ground Truth Response ></p> <p>{generations_text}</p> <p>Scoring Criteria: For each generated response, assign a score in [0, 1] based on how accurately it reflects the ground truth response.</p> <p>Guidelines:</p> <ol style="list-style-type: none"> 1. Extract 1-3 key points: <ul style="list-style-type: none"> - Extract K key points from the ground truth response along the response dimension (e.g., if evaluating a "stance", pick key points related to the stance like "clearly disagrees with X", if evaluating a "response", pick key points about the response like "offers a solution to Y"). - Because you are evaluating the full response, consider all major content, style, and intent cues expressed in the ground truth response. - Each key point should be specific and distinct. 2. Score how well the generated response matches each key point: <ul style="list-style-type: none"> - For each key point i, compare it with the generated response and assign a match value m_i in range [0, 1]: <ul style="list-style-type: none"> - 1.0: The key point is precisely and perfectly reflected. - [0.7, 0.9]: Mostly reflected with small imperfections. - [0.4, 0.6]: Partially reflected or vague, but still leaning in the correct direction. - [0.1, 0.3]: Very weak reflection. - 0.0: Missed, contradicted, or reversed. 3. Compute coverage $C = (m_1 + m_2 + \dots + m_K) / K$, which measures how comprehensive the generated response reflects the ground truth response. 4. Compute penalty P for extra or conflicting content: <ul style="list-style-type: none"> - Examine additional content in the generated response beyond those key points: <ul style="list-style-type: none"> - Does it introduce unsupported evidence and assumptions? - Is it irrelevant to what ground truth response expresses? - Is it only using generic commentary or high-level framing that misses the ground truth's goals, values, communication style, beliefs, and emotions specific to [HUMAN]? - Set a penalty P in [0, 1]: <ul style="list-style-type: none"> - 0.0: No problematic extra content; everything is perfectly matched. - [0.1, 0.3]: Slightly unnecessary, mildly speculative, or generic detail; meaning essentially unchanged. - [0.4, 0.6]: Moderate speculative, irrelevant, or vague content that somewhat shifts emphasis or adds unsupported ideas.

Figure 17: Exact HumanLM response-only similarity judge prompt used in sim evaluation, split across three parts for typesetting.

Response Similarity Judge (Part 2/3)

- [0.7, 0.9]: Significant speculative, misleading, or conflicting content that clearly changes the meaning.
- 1.0: Mostly off-topic, contradictory, or dominated by incorrect/hallucinated content.
- Penalize sycophantic openings, especially formulaic phrases such as "you're absolutely right", "totally", "completely agree", "100%

5. Response-specific checks:

- The generated response may or may not reuse phrases from the context; however, if the generated response just directly copies previous context, without quoting it, treat that as off-task behavior and give a score of 0.
- Wrong-perspective hard zero: if the generated response treats another user's perspective, identity, or experience in the thread as [HUMAN]'s own, or speaks from another participant's first-person perspective, give a score of 0.
- Assistant-like hard zero: if the ground truth is [HUMAN]'s next question, follow-up, or short request, but the generated response behaves like an assistant reply instead by directly answering the earlier prompt, giving a polished explanation, or presenting a structured helpful breakdown, treat that as wrong perspective and give a score of 0.

6. Compute the final score = $\max(0, \min(1, C - P))$

Additional considerations:

- Follow the instruction carefully.
- Be strict and reserve scores above 0.8 for clearly outstanding matches.
- Do not reward verbosity or generic topical plausibility; reward user-specific evidence.

Output format (JSON):

```
{
  "key_points": "<analysis of key points from ground truth along response dimension>",
  "1": {"thought": "<how well the 1st generated response matches each key point and compute the final score>", "score": <score>},
  "2": ...
}
```

Format Notes:

- All text in "key_points" and "thought" fields MUST be on a single line with no line breaks or newlines.
- Use standard JSON string format with double quotes. For any quotes needed inside strings, use single quotes (').
- Double check the JSON array's format, especially the comma and quotation marks.
- Ensure that ALL fields, especially "thought" and "score", are present for each item.
- You must provide exactly {num_generations} scores for the generated response(s).

Your output:

Model Output

```
{
  "key_points": "",
  "1": {"thought": "", "score": <score>},
  "2": ...
}
```

```
Response Specificity Judge (Part 1/5)

## Task

You are evaluating {num_candidates} candidate responses for grounded specificity.

A high score means the candidate response is specifically tied to this interaction
and specifically compatible with the target user, rather than being a reusable
plausible response from an average participant.
A low score means the candidate response is broadly plausible but generic, reusable
, contradictory, unsupported, or artifact-like.

## Inputs

You will receive:
1. Target user evidence
2. The current interaction context
3. {num_candidates} candidate responses

<|Target User Evidence|>
{user_history}
<|End Target User Evidence|>

<|Interaction Context|>
{context}
<|End Interaction Context|>

<|Candidate Responses|>
{candidates}
<|End Candidate Responses|>
```

Figure 18: Exact batched response specificity judge prompt used in the experiments, split across five parts for typesetting.

Response Specificity Judge (Part 2/5)

Evaluation Criteria

Score each candidate as an absolute judgment of grounded user-specificity.

A high-scoring response should:

- fit the exact current interaction
- be compatible with the target user's observed behavior
- contain a natural situated move, stance, framing, effort level, or expression pattern that makes it specific to this user/context

A low-scoring response may be fluent, coherent, topical, or plausible, but it is weakly grounded, broadly reusable, incompatible with the user/context, or artifact-like.

Core Rules

- Do not reward general quality, fluency, politeness, balance, helpfulness, verbosity, or topical plausibility by themselves.
- Do not require explicit personal facts; user evidence is background for compatibility and distinctiveness.
- Reward specificity only when it is natural, proportionate, and tied to this exact next response.
- Penalize invented user traits, unsupported personal assumptions, wrong perspective, over-explaining, over-personalizing, persona-caricature, and assistant-like responses.
- A response that is fully compatible with the user can still be generic. Do not treat compatibility alone as specificity.
- A response that smoothly advances the interaction but could be written by many plausible participants should receive only moderate scores.
- Treat the target user's own earlier turns inside the current interaction context as strong evidence for immediate stance, effort level, style, and local move.
- Penalize uncalled-for imports of personal facts, profile details, or past behavior when the current interaction does not make them natural to say.
- If the response contradicts the context, uses the wrong speaker perspective, or invents personal facts, the overall score should be low regardless of other strengths.

Score each dimension from 0.0 to 1.0.

Use the full continuous range when appropriate. The regions below define the scale.

For batched judging, score each candidate independently using the same rubric. Do not assume any candidate is ground truth.

Response Specificity Judge (Part 3/5)

Dimensions

Dimension 1: Context Specificity

How specifically does the candidate engage with the exact local interaction?

Scoring regions:

- 0.9-1.0: Tightly tied to the exact claim, question, event, decision point, constraint, disagreement, callback, or conversational detail at issue.
- 0.7-0.9: Clearly grounded in the specific local context with only minor generic framing.
- 0.5-0.7: Responds to the main local issue but misses important details, constraints, or conversational pressure.
- 0.3-0.5: Related to the broad topic but weakly grounded in this exact context.
- 0.1-0.3: Barely connected; could fit many similar interactions.
- 0.0-0.1: Off-topic, responds to the wrong issue or speaker, or ignores the local context.

High context specificity:

- Addresses the actual local claim, request, constraint, preference, decision point, joke, disagreement, or reference.
- Tracks who said what and what the candidate is responding to.
- Uses or reacts to local details without merely copying them.
- Can be implicit, terse, joking, fragmentary, or low-effort when the local function is recoverable.
- Preserves the target user's current local trajectory when the target user has already established one in the interaction.

Low context specificity:

- Gives a generic response about the broad topic.
- Responds to the wrong participant, wrong request, or wrong part of the interaction.
- Misses the key disagreement, question, constraint, or situational detail.
- Directly answers a public-context detail but replaces the target user's likely local move with a cleaner or more obvious continuation.

Response Specificity Judge (Part 4/5)

Dimension 2: User Evidence Compatibility

How compatible is the candidate with the target user's evidence without unnaturally exposing, exaggerating, or inventing that evidence?

Scoring regions:

- 0.9-1.0: Strongly compatible with the target user's evidence and current-context behavior, with no contradiction or unnatural profile display.
- 0.7-0.9: Clearly compatible with important user evidence, stance, effort level, interaction habit, or constraints.
- 0.5-0.7: Compatible but broad, weakly distinctive, incomplete, or mildly over-explicit.
- 0.3-0.5: Only weakly compatible; mostly generic, overly smooth, over-personalized, or based on a thin user signal.
- 0.1-0.3: Little meaningful compatibility with the user evidence, or the response seems to perform a profile rather than act naturally.
- 0.0-0.1: Contradicts the user evidence, invents unsupported personal facts, uses the wrong perspective, or relies on a wrong user model.

High compatibility:

- Does not contradict the user's known behavior, values, constraints, or current-context stance.
- Uses only the amount of user-specific detail that would naturally appear in this moment.
- Preserves the user's likely local move instead of replacing it with a profile-shaped explanation.
- Can reflect distinctive user evidence through form, effort level, stance, humor, skepticism, brevity, or refusal to elaborate.

Low compatibility:

- Contradicts the user evidence or current-context behavior.
- Invents personal facts, preferences, motives, or biography.
- Pulls in user history or profile details when the interaction does not call for them.
- Sounds like a persona sketch, averaged user model, or assistant summary of the user.
- Is merely non-contradictory and locally sensible, but gives no meaningful evidence of this user.

Response Specificity Judge (Part 5/5)

Overall Score

Compute the overall score with these weights:

- context_specificity: 50%
- user_evidence_compatibility: 50%

Apply these caps after computing the weighted score:

- If the response contradicts the context, uses the wrong speaker perspective, or cannot work as the next response, overall should be at most 0.20.
- If the response invents personal facts or unnaturally imports user-history/profile details, overall should be at most 0.40.

Output Requirements

- Return valid JSON only.
- Do not include Markdown.
- Each reason must be brief: one short sentence.
- Each dimension score and overall score must be a number in [0.0, 1.0].
- Use the full continuous range when appropriate; do not restrict yourself to the region boundaries.
- The overall score must follow the weighted formula and caps above.
- Return exactly one result for each label: {label_list}.

Return this JSON structure:

```
{output_schema_example}
```

Your output:

Model Output

```
{
  "label_1": {
    "context_specificity": {"score": 0.0, "reason": "..."},
    "user_evidence_compatibility": {"score": 0.0, "reason": "..."},
    "overall": 0.0,
    "reasoning": "..."}
  },
  "label_2": {
    "context_specificity": {"score": 0.0, "reason": "..."},
    "user_evidence_compatibility": {"score": 0.0, "reason": "..."},
    "overall": 0.0,
    "reasoning": "..."}
  }
}
```

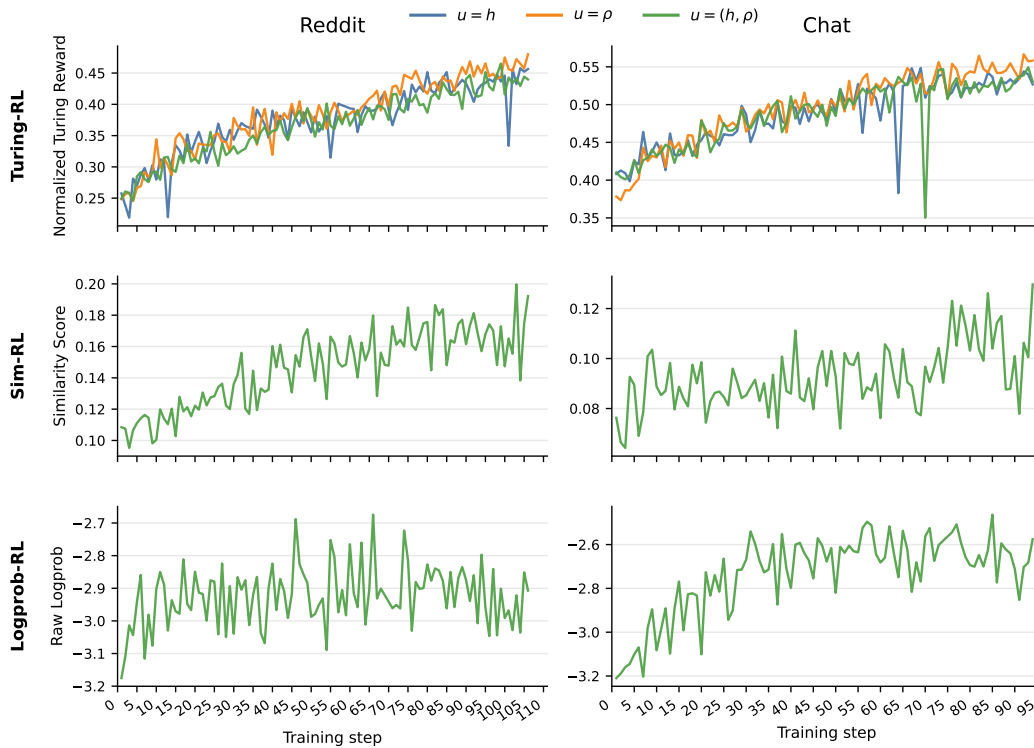


Figure 19: Raw per-step GRPO training scores for Reddit and Chat. The first row shows the unadjusted Turing reward for Turing-RL under user input ablations $u = h$, $u = \rho$, and $u = (h, \rho)$. The second and third rows show the raw similarity and log-probability reward scores for Sim-RL and Logprob-RL. Curves use every logged training step.



Figure 20: Qualitative examples from Chat and Reddit. Each column shows the conversation context, the ground truth user response, and generations from five models. GPT-5, Qwen3.5-397B-A17B, Logprob-RL, and Qwen3-8B produce verbose responses.