REINFORCING HUMAN BEHAVIOR SIMULATION VIA VERBAL FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans learn social norms and behaviors from verbal feedback (e.g., a parent saying "that was rude" or a friend explaining "here's why that hurt"). Yet, LLM learning from feedback has largely focused on domains like code and math, where rewards are verifiable and easily reduced to scalar values. As LLMs are increasingly used to simulate users, patients, students, and other personas, making them more human-like requires learning from a fundamentally different signal: feedback that is verbal, subjective, and multi-faceted. We present **DITTO**, a model trained by treating verbal feedback as a first-class signal in reinforcement learning. After each rollout, DITTO receives verbal feedback and generates a feedback-conditioned improved rollout; both outputs are jointly optimized with GRPO, distilling verbal guidance into the base policy without requiring feedback at test time. We also introduce **SOUL** (Simulation gym Of hUman-Like behavior), a unified benchmark and training data suite spanning *10 tasks* across six categories: Theory of Mind, character role play, social skill, learner simulation, user simulation, and persona simulation. DITTO achieves an average 36% improvement over the base model and exceeds GPT-5.4 on 6 of 10 SOUL benchmarks, demonstrating that RL with verbal feedback is a promising direction for training LLMs to simulate human behavior.

1 INTRODUCTION

Modeling human behavior has long been recognized as a crucial step toward imbuing artificial intelligence with social intelligence (Tomasello, 2019; Nass & Moon, 2000). From the Turing test to modern user simulators, the drive to faithfully model how people think, speak, and act has evolved into a practical necessity for training and evaluating collaborative AI systems. Today, LLMs are increasingly called upon to serve as these vital proxies—standing in for users in dialogue systems (Davidson et al., 2023; Sekulic et al., 2024), for patients in clinical settings (Schmidgall et al., 2024), for students in educational platforms (Ross & Andreas, 2025; Dou et al., 2025), and for diverse personas in social and safety testing (Zhou et al., 2024; 2025; Wang et al., 2025).

Yet current LLM simulators fall short, exhibiting systematic Sim2Real gaps—from homogeneous, overly cooperative outputs lacking individual variation (Jiang et al., 2025; Chopra et al., 2026; Jeong et al., 2026), to misaligned cognitive biases (Tjuatja et al., 2024), superhuman behaviors (Oh & Linzen, 2025; Balog & Bakken, 2026), and consistent divergences from real users in downstream tasks (Zhou et al., 2026b; Mehri et al., 2026; Chen, 2026). One reason for this gap may lie in *how* these models are trained: alignment makes models normative rather than descriptive of human behavior (Shapira et al., 2026). Reinforcement learning for LLMs, whether RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), or RL with verifiable rewards (DeepSeek-AI, 2025; Shao et al., 2024), reduces all feedback to a scalar reward or binary preference. This suffices when correctness is verifiable: a proof is right or wrong, a test passes or fails. But simulating human behavior is not about correctness; it requires behavior that is socially grounded, individually varied, and contextually appropriate. A score of 0.6 says little about *why* a simulated user was too polite or predictable, let alone how to improve it. Humans learn social behavior through verbal feedback, including explanations, corrections, and dialogue (Tomasello, 2019), and we argue that human simulators should be trained similarly.

We present **DITTO** (illustrated in Figure 1), a model that incorporates **verbal feedback** as a first-class signal in reinforcement learning for human behavior simulation (Song et al., 2026; Shi et al., 2026a;

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

Tayal et al., 2026). After each rollout, DITTO receives verbal feedback and generates an improved rollout conditioned on it; both outputs are jointly optimized via GRPO, so that the base policy internalizes the verbal guidance without requiring it at test time (§3). Crucially, our approach is flexible across task formats: it handles discriminative tasks (multiple-choice, binary classification), generative tasks (open-ended dialogue), and multi-turn interactions where the model must sustain human-like behavior across an entire conversation, all within a single unified training framework.

To train and evaluate DITTO, we introduce **SOUL** (**S**imulation gym **O**f h**U**man-**L**ike behavior; Figure 2), a unified benchmark and training data suite comprising *10 tasks* across six categories of human behavior modeling: theory of mind (Kim et al., 2023; He et al., 2024; Jin et al., 2024; Nematzadeh et al., 2018; Sclar et al., 2023), character role play (Wang et al., 2026b), social skill (Zhou et al., 2024), learner simulation (Ross & Andreas, 2025), user simulation (Hathidara et al., 2026; Naous et al., 2025), and persona simulation (Du et al., 2025). Notably, SOUL addresses a key bottleneck for RL-based human simulation: the absence of training data; most existing human simulation benchmarks provide only evaluation sets, not training splits. In SOUL, we pair every task with curated RL training data, adapting official splits where available and constructing new ones where not, such as 6,000 synthetic belief-tracking scenarios for HiToM, and 1,335 persona items for TwinVoice built from three external corpora via a two-stage LLM generation and filtering pipeline. The resulting mixture spans discriminative, generative, multi-turn, and multi-agent formats within a single unified RL training framework (§4).

Our experiments on SOUL yield several findings. First, our Qwen3-8B DITTO model achieves an average normalized score of 0.726, a 36% improvement over the Qwen3-8B base model, while matching or exceeding GPT-5.4 on 6 of 10 tasks. Gains are broad: DITTO outperforms GPT-5.4 on all three Theory of Mind tasks, on both user simulation tasks (e.g., scoring 0.930 compared to GPT-5.4's 0.575 on UserLLM), and on Sotopia social skill (0.470 vs. 0.300), spanning discriminative, generative, and multi-turn formats. Second, verbal feedback improves over standard GRPO on 8 of 10 tasks, with the largest gains on TwinVoice (+14 points), ToMi (+11), LifeChoices (+11), and UserLLM (+7). Third, analysis of training dynamics shows that verbal feedback accelerates learning—on Sotopia, DITTO outperforms GRPO from the earliest training steps—and helps mitigate safety risks: DITTO better preserves secret-keeping throughout RL optimization, whereas standard GRPO degrades on this dimension.

Contributions. (i) We show that **verbal feedback is a crucial training signal** for human behavior simulation, providing actionable, dimension-specific guidance beyond scalar rewards, improving learning speed and reducing safety risks such as secret leakage. (ii) We introduce **DITTO**, which distills verbal feedback into the base policy via feedback-conditioned rollouts jointly optimized with GRPO, achieving a 36% improvement over the base model and matching or exceeding GPT-5.4 on 6/10 tasks with an 8B model. (iii) We release **SOUL**, a unified training and evaluation environment for human-like behavior across 10 tasks in six categories, covering discriminative, generative, multi-turn, and multi-agent settings, with curated RL training data for each task.

2 RELATED WORK

2.1 INCORPORATING VERBAL FEEDBACK IN RL TRAINING

Standard RL for language models—whether RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), or GRPO (DeepSeek-AI, 2025; Shao et al., 2024)—reduces all feedback to scalar rewards or binary preferences. To handle complex feedback signals, another line of work decomposes the scalar reward into finer-grained or multi-dimensional signals while remaining in the scalar regime (Wu et al., 2023; Yu et al., 2025; Sun et al., 2025). These methods enrich the reward structure but still compress each dimension into a scalar, losing rich nature of verbal feedback.

To incorporate verbal feedback into improving LLM agents, prior work has explored supervised fine-tuning (Scheurer et al., 2023; Liu et al., 2024) or inference-time self-reflection (Shinn et al., 2023; Madaan et al., 2023), but these approaches lack the exploration capacity of RL or do not update model weights from experience. A recent line of work instead incorporates verbal feedback directly into RL optimization (Tayal et al., 2026). Self-Rewarding Language Models (Yuan et al., 2024) generate verbal self-judgments to construct preference pairs for iterative DPO. Closest to our setting are Experiential RL (Shi et al., 2026a) and the concurrent RLTF (Song et al., 2026), which likewise

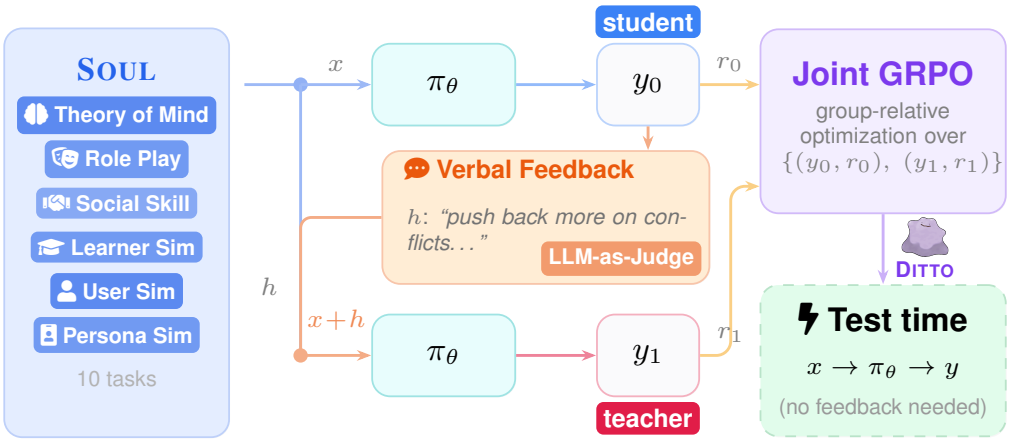


Figure 1: Overview of DITTO. ① Given prompt x from SOUL, the policy generates a student rollout y_0, scored as r_0. ② An LLM judge produces **verbal feedback**: per-dimension critiques and a reflection h. ③ h is concatenated with x and fed into π_θ to generate a teacher rollout y_1, scored as r_1. ④ Both are jointly optimized via GRPO. At test time, no feedback is needed.

internalize feedback-conditioned improvements into the policy; we detail the algorithmic differences in Section 3.4. Newer variants learn or aggregate the feedback (Singh et al., 2026; Zhang et al., 2026; Huang et al., 2026) or internalize self-critiques (Lin et al., 2026; Amin & Yildirim, 2026; Zheng et al., 2026); these works focus extensively on domains with verifiable rewards, such as mathematics and reasoning puzzles. Related work guides RL with expert traces or hints: LUFFY (Yan et al., 2025) mixes expert trajectories into GRPO, BREAD (Zhang et al., 2025b) branches rollouts from expert prefixes, and self-hinting or off-context GRPO condition rollouts on privileged hints (Liao et al., 2026; Agrawal et al., 2026). In contrast, DITTO uses judge-generated verbal feedback and self-generated teacher rollouts, requiring no expert trajectories.

2.2 SIMULATING HUMAN BEHAVIOR WITH LLMS

LLM-based human simulation spans a broad landscape of tasks. User simulation has evolved from rule-based and statistical approaches for dialogue management (Schatzmann et al., 2006; Li et al., 2016; Ni et al., 2026) to LLM-based simulators for task-oriented dialogue (Davidson et al., 2023; Sekulic et al., 2024; Sun et al., 2022), information-seeking (Sun et al., 2021; Zhang et al., 2024; Ren et al., 2024), and recommendation (Wang et al., 2025; Bougie & Watanabe, 2025). Beyond user simulation, LLMs are increasingly used for social interaction (Zhou et al., 2024; 2025), character role-play (Wang et al., 2026b; Tang et al., 2026; Shi et al., 2026b), persona simulation (Samuel et al., 2025), and learner simulation (Ross & Andreas, 2025).

However, recent efforts have revealed systematic *Sim2Real gaps*: LLM-simulated human behavior shows systematic differences compared to real humans (Seshadri et al., 2026; Tjuatja et al., 2024; Zhou et al., 2026b; Mehri et al., 2026; Shapira et al., 2026), which undermine the utility of simulators for downstream evaluation and training. Training-based approaches have begun to address this—UserLLM (Naous et al., 2025) via SFT on human conversations, Turing-RL (Wang et al., 2026c) via an indistinguishability reward, and Sotopia-RL (Yu et al., 2025) via RL with multi-dimensional rewards (see also Wang et al., 2026a; Feng et al., 2026). However, each targets a narrow set of tasks (e.g., Yu et al. (2025) focus only on social skills, Naous et al. (2025) focus only on user simulation). Our work incorporates verbal feedback into RL to improve a single model across the 10 diverse tasks in SOUL, spanning six categories of human behavior modeling.

3 METHOD

3.1 PROBLEM FORMULATION

We study the problem of training a language model for human simulation tasks. Given an input x (e.g., an instruction, scenario, dialogue history, or user profile), the model generates an output

trajectory

$$y \sim \pi_\theta(\cdot \mid x),$$

where π_θ is the current policy.

Each sampled trajectory is evaluated by a task-specific judge or simulator, which returns supervision for learning. In the standard RL setting, this supervision is reduced to a scalar reward

$$r(x, y) \in \mathbb{R}.$$

Our goal is to train the policy π_θ to maximize reward in these human simulation tasks, while allowing training to exploit richer judge supervision when available.

3.2 Verbal Feedback

In many human simulation tasks, the judge provides not only a scalar reward but also rich verbal feedback, such as dimension-wise scores, textual critiques, and task-specific suggestions. Formally, for a rollout $y \sim \pi_\theta(\cdot \mid x)$, the judge returns

$$(r, h) = \mathcal{J}(x, y), \tag{1}$$

where r is a scalar reward and h is verbal feedback (per-dimension critiques and a natural-language reflection) describing what was done well, what went wrong, and how the behavior can be improved.

This feedback is often highly informative. For example, in human behavior simulation, it may reflect dimensions such as believability, relationship maintenance, knowledge consistency, goal achievement, and adherence to social norms. Standard RL methods typically compress all such information into a single scalar reward, which can lose useful learning signal.

To ensure robustness of learning, we cast h as privileged information (Vapnik & Vashist, 2009; Vapnik & Izmailov, 2015): it is available during training but not at test time. Our goal is therefore to use this verbal feedback to improve the base policy $\pi_\theta(\cdot \mid x)$, while keeping inference-time generation unchanged.

3.3 Learning from Verbal Feedback

We aim to optimize the base policy $\pi_\theta(y \mid x)$ using reward and feedback during training, so the policy internalizes feedback-driven improvements without requiring feedback at inference time.

To achieve this, for each sampled rollout

$$y_0 \sim \pi_\theta(\cdot \mid x), \tag{2}$$

we ask the judge to provide reward and feedback

$$(r_0, h) = \mathcal{J}(x, y_0), \tag{3}$$

and then sample a second rollout conditioned on the same feedback:

$$y_1 \sim \pi_\theta(\cdot \mid x, h), \qquad r_1 = R(x, y_1). \tag{4}$$

Here, y_0 and y_1 are produced by the *same policy with shared weights* and differ only in their prompting context: y_1 additionally sees the verbal feedback h. Following the self-distillation literature, we call y_0 the *student* rollout and y_1 the *teacher* rollout.

We optimize both behaviors during training: the base policy learns from feedback-induced improvements without conditioning on feedback at inference, and an additional GRPO objective on the feedback-conditioned rollouts makes the model better at acting on feedback.

Concretely, for each prompt x, we first sample G original rollouts $\{y_{i,0}\}_{i=1}^{G}$ from the base policy. For each $y_{i,0}$, we obtain judge feedback h_i and generate a corresponding feedback-conditioned rollout $y_{i,1}$. This gives the training group

$$\mathcal{G}(x) = \{y_{i,0}, y_{i,1}\}_{i=1}^{G}, \tag{5}$$

which contains both the original rollouts and their feedback-conditioned counterparts.

We compute group-relative advantages over all samples in $\mathcal{G}(x)$:

$$A_{i,j} = \frac{r_{i,j} - \mu_r}{\sigma_r}, \qquad \mu_r = \frac{1}{2G}\sum_{i=1}^{G}\sum_{j=0}^{1} r_{i,j}, \qquad \sigma_r = \mathrm{std}(\{r_{i,j}\}_{i=1,j=0}^{G,1}). \tag{6}$$

We then apply the standard clipped GRPO objective to this group:

$$\mathcal{L}_{\text{group}} = -\mathbb{E}_x\left[\sum_{i=1}^{G}\sum_{j\in\{0,1\}}\sum_{t=1}^{|y_{i,j}|}\min\Big(\rho_{i,j,t}A_{i,j}, \mathrm{clip}(\rho_{i,j,t}, 1-\epsilon, 1+\epsilon)A_{i,j}\Big)\right], \tag{7}$$

where

$$\rho_{i,j,t} = \frac{\pi_\theta(y_{i,j,t} \mid x, y_{i,j,<t})}{\pi_{\text{old}}(y_{i,j,t} \mid x, y_{i,j,<t})}. \tag{8}$$

In addition, we apply an extra GRPO update on the feedback-conditioned rollouts $\{y_{i,1}\}_{i=1}^{G}$ alone, treating them as a separate GRPO group:

$$\mathcal{L}_{\text{fb}} = -\mathbb{E}_x\left[\sum_{i=1}^{G}\sum_{t=1}^{|y_{i,1}|}\min\Big(\hat{\rho}_{i,t}\hat{A}_{i,1}, \mathrm{clip}(\hat{\rho}_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,1}\Big)\right], \tag{9}$$

where $\hat{A}_{i,1}$ is computed by normalizing rewards within the feedback-conditioned group $\{r_{i,1}\}_{i=1}^{G}$, and $\hat{\rho}_{i,t} = \pi_\theta(y_{i,1,t} \mid x, h_i, y_{i,1,<t})/\pi_{\text{old}}(y_{i,1,t} \mid x, h_i, y_{i,1,<t})$ is the importance ratio under the feedback-conditioned context. The final training loss is

$$\mathcal{L} = \mathcal{L}_{\text{group}} + \mathcal{L}_{\text{fb}}. \tag{10}$$

$\mathcal{L}$ trains on three (rollout, context) pairs: (y_0, x) and (y_1, x) in $\mathcal{L}_{\text{group}}$, and $(y_1, x{+}h)$ in $\mathcal{L}_{\text{fb}}$. The (y_1, x) term distills feedback into the base policy by optimizing feedback-improved rollouts without exposing h. We normalize $\mathcal{L}_{\text{fb}}$ separately because x and $x{+}h$ induce different conditional distributions and reward baselines; pooling them would bias the advantages.

3.4 Connections to Other RL Approaches

Our approach is related to several recent works on RL with textual feedback:

- **RLTF.** RL from Text Feedback (RLTF; Song et al., 2026) similarly generates a student output followed by a feedback-conditioned teacher output and trains the policy to internalize the improvement. Our method differs by (1) using GRPO over the joint group y_0, y_1 rather than AWR (Peng et al., 2019), (2) conditioning the teacher on a judge-generated reflection h rather than raw feedback, and (3) targeting social simulation with noisy, subjective, and multi-dimensional feedback, rather than tasks with *verifiable rewards* as in RLTF-SD. See section 6.2 for comparison.
- **ERL.** Experiential RL (ERL) (Shi et al., 2026a) uses SFT on reject-sampled teacher rollouts to internalize feedback. In contrast, we train on all teacher rollouts with an RL objective, differentiating rollout quality under continuous rewards and improving sample efficiency. We compare against an ERL variant in our ablation study.
- **SDPO.** Self-distillation policy optimization methods distill textual feedback with a reverse KL objective (Vapnik & Vashist, 2009; Vapnik & Izmailov, 2015; Shenfeld et al., 2026; Ye et al., 2026b; Hubotter et al., 2026). Our approach instead samples from the teacher and optimizes with group-normalized advantage. We further optimize the feedback-conditioned model via $\mathcal{L}_{\text{fb}}$ rather than using an EMA model. LLM-as-a-Coach (Ye et al., 2026a) applies this KL-based recipe to judge feedback on open-ended tasks, and recent analyses show that the privileged self-teacher can collapse the student's output distribution (Robert & Qader, 2026; Peng et al., 2026; Yang et al., 2026). We compare against SDPO-like variants in our ablation study.

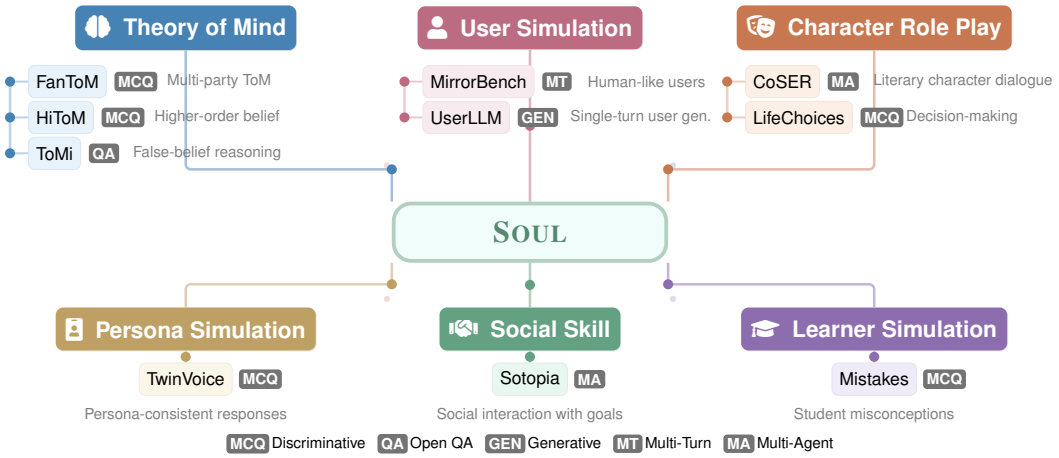


Figure 2: Overview of SOUL, our unified benchmark and training data suite spanning 10 tasks across six categories of human simulation.

4 SOUL: SIMULATION GYM OF HUMAN-LIKE BEHAVIOR

A key bottleneck in training and evaluating human simulators is the lack of a unified resource covering the full breadth of human-like behavior. Existing benchmarks each focus on a narrow slice, such as Theory of Mind, role-play, social interaction, or user simulation, and use incompatible formats, metrics, and data splits (Yu et al., 2025; Du et al., 2026; Wang et al., 2026b; Wu et al., 2026). To address this, we introduce **SOUL**-Bench (**S**imulation gym **O**f h**U**man-**L**ike behavior), a unified benchmark and training suite spanning 10 tasks across six categories of human simulation (Figure 2). For each task, SOUL provides a *held-out test set* for evaluation and a *curated training set* for RL. When training splits are available, we use them directly; otherwise, we construct training contexts from held-out evaluation data or related datasets. Full task descriptions, examples, and statistics are provided in Section D.

- **Theory of Mind.** **FanToM** (Kim et al., 2023) tests ToM in multi-party conversations with information asymmetry via MCQ and open-ended questions about who knows what (1,086 test; 18 held-out conversations for evaluation, the rest for training). **HiToM** (He et al., 2024) tests higher-order nested beliefs with deception and private vs. public communication (100 test; MCQ; 200 generated stories giving 6,000 training examples). **ToMi** (Nematzadeh et al., 2018) is a classic false-belief test, used in its paraphrased ParaToMi form (Sclar et al., 2023) to test ToM beyond surface cues (600 test; QA; 6,004 training examples).
- **Character Role Play. CoSER** (Wang et al., 2026b) covers literary characters from 771 books: given a scenario, character profile, and motivations, the model conducts a multi-turn dialogue (up to 20 turns), scored by an LLM judge on storyline consistency, anthropomorphism, character fidelity, and storyline quality; we evaluate 200 conversations (100 ID, 100 OOD) and train on its 26.5k dialogues. **LifeChoices** (Xu et al., 2024a) tests persona-driven decision making on 1,462 decision points from 388 novels: given context, scenario, and identity, the model selects the option matching the character's original decision, emphasizing behavior-level fidelity and reasoning over the character's knowledge and evolving motivations.
- **Social Skill. Sotopia** (Zhou et al., 2024) evaluates social skills via dyadic simulation: two agents with private social goals interact for up to 20 turns, and an LLM judge scores the interaction on seven dimensions (believability, relationship, knowledge, secret-keeping, social rules, financial outcomes, goal achievement). Following Zhou et al. (2026a), we evaluate on the 100-scenario Sotopia-hard split and train on the 2,310 non-overlapping Sotopia-π scenarios (Yu et al., 2025).
- **Learner Simulation. Mistakes** (Ross & Andreas, 2025; Do et al., 2026) tests whether models can simulate *incorrect* student reasoning in K-12 math: given a multiple-choice problem and a misconception description, the model must pick the wrong answer a student with that misconception would choose (600 test; accuracy). Training instances come from the Eedi "Mining Misconceptions in Mathematics" dataset, one per labeled (question, wrong-option) pair.

- **User Simulation. MirrorBench** (Hathidara et al., 2026) evaluates multi-turn user simulation: given a synthesized user goal, the model generates every user turn while an LLM replays the assistant, and the result is compared with the original human dialogue via lexical diversity (MATTR, HD-D, Yule's K) and LLM-judge realism scores (795 test episodes; ∼3,400 training conversations). **UserLLM** (Naous et al., 2025) evaluates single-turn user simulation: given context and intent, the model generates one user message, evaluated on CSQA (role adherence), NQ (intent adherence), and PRISM (diversity, intent decomposition, termination prediction, AI-detection; 1,800 test); training data are built from PRISM, NaturalQuestions, and CommonsenseQA.
- **Persona Simulation. TwinVoice** (Du et al., 2025) is a discriminative persona test where, given a user's conversation history and an anchor post, the model selects the response matching the user's communication style. Its 5,687 instances cover social, interpersonal, and narrative personas and are evaluated by accuracy. For training we construct 1,335 instances from Bluesky, Pushshift Telegram, and Project Gutenberg, with GPT-5.4-mini generating candidates from persona histories and GPT-5.4 filtering them.

5 Experimental Setup

Implementation Details. We use Qwen3-VL-8B-Instruct[1] as the base model. We sample 1,024 instances from each SOUL task training split, for a total of 10,240 training instances. Training is implemented in Verl with LoRA, a batch size of 64, a group size of 8, and around 100 steps; the same recipe is used for all methods and tasks (full hyperparameters and the recipe selection in Section B). Across our experiments, we use `gpt-5-nano` as the judge model and environment model, except for CoSER, where we use `gpt-5.4` as the judge model to provide more stable evaluations.

Baselines. We compare against (1) proprietary models GPT-5.4 and GPT-5-nano, the latter also providing training feedback; (2) open-source models specialized for human simulation—Sotopia-RL-7B (Yu et al., 2025), HER-32B (Du et al., 2026), Coser-8B (Wang et al., 2026b), and HumanLM-8B (Wu et al., 2026)—evaluated on our tasks with the best result reported per task; and (3) Base, the base model Qwen3-VL-8B-Instruct, and GRPO, the base model trained with GRPO under the same setup and data.

Evaluation Protocol. We evaluate each model on all 10 SOUL tasks using the metrics defined in Section 4, sampling 100 instances per task given the size of many test sets, and report the average score for tasks with multiple metrics. We additionally conduct human annotation on the four judge-scored SOUL benchmarks: 25 randomly sampled tasks each from CoSER, MirrorBench, and Sotopia and 15 from UserLLM (NQ), giving 90 tasks and 270 conversations simulated by GPT-5.4, Qwen3-VL-8B-Instruct, and DITTO-8B (results in section 6.3). The three outputs for a task were assigned to the *same* Prolific annotator in random order, controlling for position and between-subject effects, and rated with the benchmark-specific rubric (e.g., Sotopia's seven dimensions); see §B.1 for annotator demographics and the interface.

6 Experimental Results

6.1 Main Results

Table 1 presents results across SOUL tasks. We make the following observations: (1) Our trained models outperform proprietary and specialized baselines, demonstrating the effectiveness of SOUL for improving human-simulation capabilities. Both GRPO and DITTO achieve higher average scores than the strongest baseline, GPT-5.4. DITTO outperforms GPT-5.4 on 6 of 10 tasks and all previous specialized human-simulation models across all tasks. (2) Compared with GRPO, DITTO achieves an average absolute improvement of 4.8(3) Although we use GPT-5-nano to provide feedback during training, DITTO consistently outperforms GPT-5-nano across all tasks. This suggests that the gains come from learning from its feedback (e.g., multi-dimensional evaluation and error analysis), rather than simply distilling a stronger model. Moreover, six tasks (FanToM, HiToM, ToMi, LifeChoices, Mistakes, TwinVoice) use exact-match accuracy without an LLM judge, and DITTO improves over GRPO on five of them, indicating that the gains are not an artifact of judge preference. (4) *Task-wise*,

[1] https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct

Category	Benchmark	GPT-5.4	GPT-5-nano	Others	Base	GRPO	DITTO
ToM	FanToM	0.900	0.720	0.780[a]	0.780	0.940	**0.950**
	HiToM	0.700	0.370	0.560[c]	0.580	0.770	**0.780**
	ToMi	0.880	0.850	0.720[a]	0.680	0.820	**0.930**
Role Play	CoSER	**0.659**	0.352	0.419[c]	0.435	0.541	0.512
	LifeChoices	**0.870**	0.600	0.750[c]	0.670	0.690	0.800
Social Skill	Sotopia	0.300	0.310	0.312[b]	0.277	0.423	**0.470**
Learner Sim.	Mistakes	0.570	**0.580**	0.520[a]	0.460	**0.580**	0.560
User Sim.	MirrorBench	0.536	0.358	0.481[a]	0.547	0.683	**0.713**
	UserLLM	0.575	0.324	0.537[c]	0.469	0.863	**0.930**
Persona Sim.	TwinVoice	**0.640**	0.230	0.390[c]	0.430	0.470	0.610
Average		0.663	0.469	0.547	0.533	0.678	**0.726**

Table 1: **Main results on SOUL tasks.** We report the primary metric for each benchmark (higher is better). *Others* refers to the best result achieved by other specialized human-simulation models, including [a]HumanLM-8B (Wu et al., 2026), [b]Sotopia-RL-7B (Yu et al., 2025), [c]HER-32B (Du et al., 2026), [d]Coser-8B (Wang et al., 2026b). *Base* denotes the base model, Qwen3-VL-8B-Instruct. **Bold** indicates the best result in each row.

DITTO shows larger gains on several tasks where verbal feedback provides richer behavioral guidance (e.g., Sotopia, UserLLM, LifeChoices, and TwinVoice). Gains are smaller on FanToM and HiToM, while DITTO is slightly below GRPO on CoSER and Mistakes, suggesting that the benefit of verbal feedback varies with the nature of the task and its evaluation criteria.

6.2 ABLATION STUDY

Figure 3 shows an ablation study on the learning algorithm. Specifically, we compare DITTO with the following methods: (1) GRPO; (2) Ditto only (y_1, x), a variant that only trains on (y_1, x) pairs and removes the loss on (y_0, x) and $(y_1, x{+}h)$ pairs; (3) Ditto no $\mathcal{L}_{\text{fb}}$, which removes $\mathcal{L}_{\text{fb}}$, i.e., the $(y_1, x{+}h)$ term, and keeps only $\mathcal{L}_{\text{group}}$; (4) RLTF-SD (Song et al., 2026), which use a specific AWR version for advantage calculation; (5) ERL (Shi et al., 2026a), which uses a supervised fine-tuning objective on y_1 instead of RL; and several SDPO (Hübotter et al., 2026) variants including (6) SDPO+ token, which combines SDPO with token-level loss and the GRPO objective; (7) SDPO+ logits, which combines SDPO with logits-level loss and the GRPO objective; (8) SDPO token, i.e., SDPO with token-level loss; and (9) SDPO logits, i.e., SDPO with logits-level loss.

From the results, DITTO outperforms GRPO on most sub-metrics, particularly auxiliary metrics such as *secret*, which measures whether the agent avoids leaking private information. This safety-critical dimension is not directly optimized by the scalar RL reward but can be explicitly addressed through verbal feedback. While GRPO collapses multi-dimensional scores into a scalar, losing fine-grained learning signals, DITTO can learn from feedback across individual dimensions. Among other feedback-based methods, reverse-KL methods such as SDPO collapse on most metrics except *secret*, consistent with recent reports of collapse in on-policy self-distillation (Robert & Qader, 2026; Peng et al., 2026). Combining SDPO with GRPO improves performance but still underperforms DITTO. ERL also underperforms, likely because the lack of reward normalization leads to unstable learning under noisy feedback. See Appendix C for additional results. DITTO is also stable across random seeds (4.69 ± 0.02 over four Sotopia runs) and transfers to a smaller Qwen3-4B base (4.51, above 8B GRPO at 4.23); see Appendix B.

Teacher-Student Gap Analysis Figure 8 plots the score gap between the teacher and student responses on Sotopia. The teacher consistently outperforms the student, and the gap widens over training, i.e., verbal feedback keeps providing an informative learning target. DITTO maintains a larger and more stable gap than the ablations: without $\mathcal{L}_{\text{fb}}$, the gap rises early but then declines, as the teacher stops improving over the student without explicitly learning from feedback—matching its training curve, which improves early but saturates later.

Qualitative Feedback Analysis Figure 5 shows feedback on one Sotopia scenario at steps 10 and 100. The feedback tracks the model's current failure modes: early on it flags basic negotiation

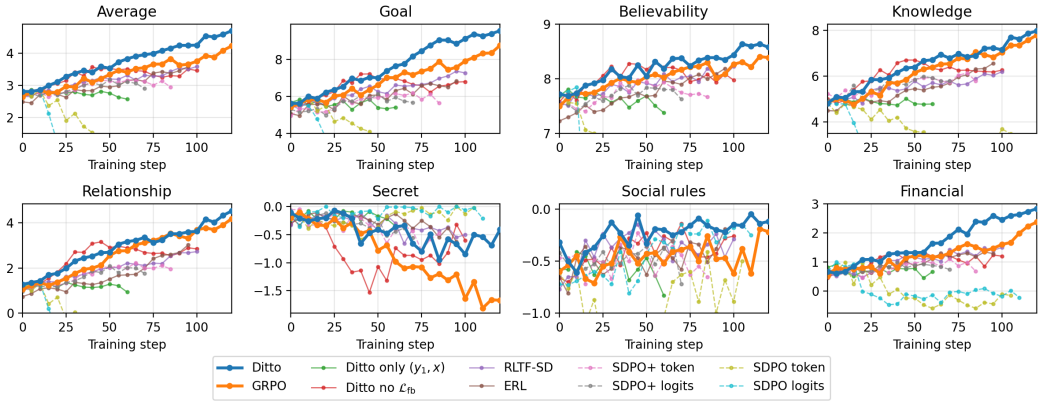


Figure 3: Ablation on Sotopia: held-out judge score (average and the seven dimensions) vs. training step for DITTO, GRPO, and eight feedback-based variants (Section 6.2); higher is better.

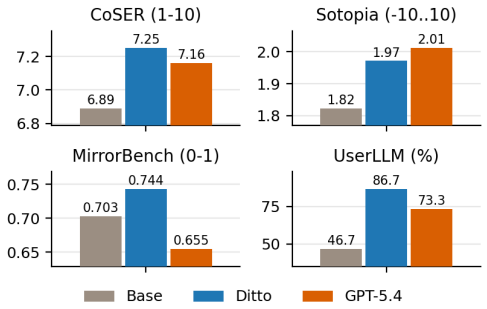


Figure 4: Human evaluation on the four judge-scored SOUL tasks (per-dim scores in Figure 7).

	Step 10	Step 100
Goal	1 no path to the $17.3 floor	→ 3 ignores his "cash-only" rule
Relationship	−1 adversarial, dismissive	→ 1 show appreciation; simplify deal
Knowledge	2 missed payment options	→ 6 unclear: price or trust?
Believability	7 theatrical, repetitive	→ 7 over-engineered bonuses

Figure 5: Verbal feedback on one Sotopia scenario at training steps 10 and 100 (circles: judge scores; full feedback in Table 3).

issues (repeatedly rejecting low offers, never steering the buyer toward a concrete commitment, dismissive language); once the model adopts more structured strategies, it shifts to finer-grained issues (over-complicated bonus or credit schemes, unclear buyer constraints, no clean closing offer). Verbal feedback thus provides diagnostic supervision beyond scalar rewards by explaining how the response should be improved.

6.3 HUMAN EVALUATION

Only four SOUL tasks are scored by an LLM judge—the other six use verifiable exact-match metrics—so we collect human judgments on exactly those four. Figure 4 reports the results. Human annotators rank DITTO above the base model on all four tasks, confirming that the gains observed under automatic evaluation carry over to human judgments. DITTO is also ranked above GPT-5.4 on three of the four: CoSER (7.25 vs. 7.16), MirrorBench (0.744 vs. 0.655), and UserLLM (86.7% vs. 73.3%). On Sotopia, DITTO is marginally below GPT-5.4 overall (1.97 vs. 2.01); it scores higher on relationship (1.80 vs. 0.56) and knowledge (5.16 vs. 4.16) but lower on believability (6.48 vs. 8.00).

7 CONCLUSION

We introduce DITTO, a reinforcement learning approach that treats verbal feedback as a first-class training signal for human behavior simulation. By generating feedback-conditioned improved rollouts and jointly optimizing them with the original outputs, DITTO enables the base policy to internalize rich, actionable guidance without requiring feedback at test time. To support this setting, we build SOUL, a unified training and evaluation suite spanning 10 tasks across six categories of human simulation. Experiments show that verbal feedback substantially improves performance over standard GRPO on 8 of 10 tasks, and human annotators prefer DITTO to GPT-5.4 on three of four tasks, while also improving training efficiency and safety-related behaviors. Together, our results suggest that moving beyond scalar rewards toward richer feedback is a promising direction for training more human-like simulators.

AI USE STATEMENT

In this work, we used generative AI tools to polish human-written drafts and check the validity of mathematical formulations. Additionally, we used generative AI tools to create analysis code and perform certain scaled evaluations. We reviewed all AI-assisted work, and LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Priyank Agrawal, Ankur Samanta, Shervin Ghasemlou, Boris Vidolov, Jalaj Bhandari, Kavosh Asadi, Daniel Jiang, and Aditya Modi. Off-Context GRPO: Learning to reason on hard problems using privileged information. *arXiv preprint arXiv:2607.19313*, 2026. URL https://arxiv.org/abs/2607.19313.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 475–484, 2019.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4473–4484, 2021.

Muhammad Zain Amin and Kibele Sebnem Yildirim. Self-Review reinforcement learning (SRRL) with Cross-Episode memory and policy distillation. *arXiv preprint arXiv:2607.05541*, 2026. URL https://arxiv.org/abs/2607.05541.

Krisztian Balog and Arild Michel Bakken. "act like a 5th grader" is not enough: Bounding knowledge in LLM-Based user simulators. *arXiv preprint arXiv:2608.30033*, 2026. URL https://arxiv.org/abs/2608.30033.

Nicolas Bougie and Narimasa Watanabe. Simuser: Simulating user behavior with large language models for recommender system evaluation. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://aclanthology.org/2025.acl-industry.5/.

Liang Chen. Simulated customers never walk away: Decision fidelity of LLM user simulators measured against real purchase outcomes. *arXiv preprint arXiv:2606.20708*, 2026. URL https://arxiv.org/abs/2606.20708.

Harshita Chopra, Kshitish Ghate, Aylin Caliskan, Tadayoshi Kohno, Chirag Shah, and Natasha Jaques. Beyond cooperative simulators: Generating realistic user personas for robust evaluation of LLM agents. *arXiv preprint arXiv:2605.12894*, 2026. URL https://arxiv.org/abs/2605.12894.

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. *arXiv preprint arXiv:2309.13233*, 2023. URL https://arxiv.org/abs/2309.13233.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Heejin Do, Shashank Sonkar, and Mrinmaya Sachan. Simulating students or sycophantic problem solving? on misconception faithfulness of LLM simulators. *arXiv preprint arXiv:2605.12748*, 2026. URL https://arxiv.org/abs/2605.12748.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. Simulatorarena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In *Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://arxiv.org/abs/2510.05444.

Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, and Yiqun Liu. TwinVoice: A multi-dimensional benchmark towards digital twins via LLM persona simulation. *arXiv preprint arXiv:2510.25536*, 2025. URL https://arxiv.org/abs/2510.25536.

Chengyu Du, Xintao Wang, Aili Chen, Weiyuan Li, Rui Xu, Junteng Liu, Zishan Huang, Rong Tian, Zijun Sun, Yuhao Li, Liheng Feng, Deming Ding, Pengyu Zhao, and Yanghua Xiao. Her: Human-like reasoning and reinforcement learning for llm role-playing. *ArXiv*, abs/2601.21459, 2026.

Xiachong Feng, Yi Jiang, Xiaocheng Feng, Deyi Yin, Libo Qin, Yangfan Ye, Lei Huang, Weitao Ma, Yuxuan Gu, Chonghan Qin, Bing Qin, and Lingpeng Kong. SAVOIR: Learning social Savoir-Faire via shapley-based reward attribution. *arXiv preprint arXiv:2604.18982*, 2026. URL https://arxiv.org/abs/2604.18982.

Ashutosh Hathidara, Julien Yu, Vaishali Senthil, Sebastian Schreiber, and Anil Babu Ankisettipalli. MirrorBench: A benchmark to evaluate conversational user-proxy agents for human-likeness. *arXiv preprint arXiv:2601.08118*, 2026. URL https://arxiv.org/abs/2601.08118.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. HI-TOM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*, 2024. URL https://arxiv.org/abs/2310.16755.

Lei Huang, Xiang Cheng, Chenxiao Zhao, Guobin Shen, Junjie Yang, Xiaocheng Feng, Yuxuan Gu, Xing Yu, and Bing Qin. Bootstrapping exploration with Group-Level natural language feedback in reinforcement learning. *arXiv preprint arXiv:2603.04597*, 2026. URL https://arxiv.org/abs/2603.04597.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. URL https://arxiv.org/abs/2601.20802.

Jonas Hubotter, Frederike Lubeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *ArXiv*, abs/2601.20802, 2026.

Daeheon Jeong, Yoonjoo Lee, Eugene Choi, Sinie van der Ben, and Juho Kim. Investigating assistant bias in LLM user simulators using a role vector. *arXiv preprint arXiv:2609.00608*, 2026. URL https://arxiv.org/abs/2609.00608.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *arXiv preprint arXiv:2510.22954*, 2025. URL https://arxiv.org/abs/2510.22954.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2401.08743.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2310.15421.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, et al. OpenAssistant conversations – democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, 2023.

Xiujun Li, Zachary C. Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*, 2016. URL https://arxiv.org/abs/1612.05688.

Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, and Jiang Bian. Self-Hinting language models enhance reinforcement learning. *arXiv preprint arXiv:2602.03143*, 2026. URL https://arxiv.org/abs/2602.03143.

Jianbo Lin, Xiaomin Yu, Yi Xin, Yifu Guo, Zhuosong Jiang, Zhongqi Yue, Weishi Wang, Heqing Zou, Chengwei Qin, and Hui Xiong. ICRL: Learning to internalize Self-Critique with reinforcement learning. *arXiv preprint arXiv:2605.15224*, 2026. URL https://arxiv.org/abs/2605.15224.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2302.02676.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.17651.

Shuhaib Mehri, Philippe Laban, Sumuk Shashidhar, Marwa Abdulhai, Sergey Levine, Michel Galley, and Dilek Hakkani-Tür. Measuring and mitigating the distributional gap between real and simulated user behaviors. *arXiv preprint arXiv:2605.07847*, 2026. URL https://arxiv.org/abs/2605.07847.

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

Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. Evaluating theory of mind in question answering, 2018. URL https://arxiv.org/abs/1808.09352.

Bo Ni, Leyao Wang, Yu Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Leura, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, Zhengzhong Tu, Alexa Siu, Zichao Wang, David 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. *arXiv preprint arXiv:2604.24977*, 2026. URL https://arxiv.org/abs/2604.24977.

Byung-Doh Oh and Tal Linzen. To model human linguistic prediction, make LLMs less superhuman. *arXiv preprint arXiv:2510.05141*, 2025. URL https://arxiv.org/abs/2510.05141.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 2022. URL https://arxiv.org/abs/2203.02155.

Keqin Peng, Chen Li, Yuanxin Ouyang, Yancheng Yuan, and Liang Ding. Diagnosing and mitigating thinking collapse in On-Policy Self-Distillation. *arXiv preprint arXiv:2607.10805*, 2026. URL https://arxiv.org/abs/2607.10805.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *ArXiv*, abs/1910.00177, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://arxiv.org/abs/2305.18290.

Ruiyang Ren, Peng Qiu, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Bases: Large-scale web search user simulation with large language model based agents. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. URL https://arxiv.org/abs/2402.17505.

Justin Robert and Raheel Qader. One symptom, three levers: A critical review of On-Policy Self-Distillation. *arXiv preprint arXiv:2608.25936*, 2026. URL https://arxiv.org/abs/2608.25936.

Alexis Ross and Jacob Andreas. Learning to make MISTAKEs: Modeling incorrect student thinking and key errors. *arXiv preprint arXiv:2510.11502*, 2025. URL https://arxiv.org/abs/2510.11502.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025. URL https://arxiv.org/abs/2407.18416.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(2):97–126, 2006. URL https://doi.org/10.1017/S0269888906000944.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2303.16755.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments, 2024. URL https://arxiv.org/abs/2405.07960.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*, 2023. URL https://arxiv.org/abs/2306.00924.

Ivan Sekulic, Silvia Terragni, Victor Guimaraes, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. Reliable llm-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT)*, 2024. URL https://arxiv.org/abs/2402.13374.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Eilam Shapira, Moshe Tennenholtz, and Roi Reichart. Alignment makes language models normative, not descriptive. *arXiv preprint arXiv:2603.17218*, 2026. URL https://arxiv.org/abs/2603.17218.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *ArXiv*, abs/2601.19897, 2026.

Taiwei Shi, Sihao Chen, Bowen Jiang, Linxin Song, Longqi Yang, and Jieyu Zhao. Experiential reinforcement learning. *arXiv preprint arXiv:2602.13949*, 2026a. URL https://arxiv.org/abs/2602.13949.

Wenlong Shi, Jianxun Lian, Mingqi Wu, Haiming Qin, Mingyang Zhou, Xing Xie, Naipeng Chao, and Hao Liao. PersonaArena: Dynamic simulation for evaluating and enhancing Persona-Level Role-Playing in large language models. *arXiv preprint arXiv:2605.17044*, 2026b. URL https://arxiv.org/abs/2605.17044.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Utsav Singh, Sidhaarth Sredharan, Souradip Chakraborty, and Amrit Singh Bedi. RL with learnable textual feedback: A bilevel approach. *arXiv preprint arXiv:2605.24547*, 2026. URL https://arxiv.org/abs/2605.24547.

Yuda Song, Lili Chen, Fahim Tajwar, Rémi Munos, Deepak Pathak, J. Andrew Bagnell, Aarti Singh, and Andrea Zanette. Expanding the capabilities of reinforcement learning via text feedback. *arXiv preprint arXiv:2602.02482*, 2026. URL https://arxiv.org/abs/2602.02482.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. URL https://dl.acm.org/doi/10.1145/3404835.3463241.

Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, M. de Rijke, and Zhaochun Ren. Metaphorical user simulators for evaluating task-oriented dialogue systems. *ACM Transactions on Information Systems*, 42:1 – 29, 2022.

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

Yihong Tang, Kehai Chen, Liang Yue, Benyou Wang, and Min Zhang. CRPO: Character-centric group relative policy optimization for role-aware reasoning in role-playing agents. *arXiv preprint arXiv:2605.25511*, 2026. URL https://arxiv.org/abs/2605.25511.

Kshitij Tayal, Arun Sharma, Genta Indra Winata, Anirban Das, and Sambit Sahu. The rise of verbal reinforcement learning. *arXiv preprint arXiv:2609.01597*, 2026. URL https://arxiv.org/abs/2609.01597.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Michael Tomasello. *Becoming Human: A Theory of Ontogeny*. Harvard University Press, Cambridge, MA, 2019. ISBN 9780674248281.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Vladimir Naumovich Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16:2023–2049, 2015.

Jianing Wang, Xintao Wang, Aili Chen, Jie Shi, Hongcheng Guo, Jun Gao, Wenxuan Zhao, Chengkun Lang, Yuanli Guo, and Yanghua Xiao. SocialRL: Refining LLMs' social intelligence through multi-turn reinforcement learning and reward design. *arXiv preprint arXiv:2609.09764*, 2026a. URL https://arxiv.org/abs/2609.09764.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji rong Wen. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 2025. URL https://dl.acm.org/doi/10.1145/3708985.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. Coser: A comprehensive literary dataset and framework for training and evaluating llm role-playing and persona simulation, 2026b. URL https://arxiv.org/abs/2502.09082.

Yingshan Susan Wang, Cedegao E. Zhang, Linlu Qiu, Zexue He, Pengyuan Li, Alex Pentland, Roger P. Levy, and Yoon Kim. Learning user simulators with turing rewards. *arXiv preprint arXiv:2606.19336*, 2026c. URL https://arxiv.org/abs/2606.19336.

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

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.01693.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. Character is destiny: Can role-playing language agents make persona-driven decisions? 2024a.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. Character is destiny: Can role-playing language agents make persona-driven decisions? *arXiv preprint arXiv:2404.12138*, 2024b. URL https://arxiv.org/abs/2404.12138.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.14945.

Fan Yang, Rui Meng, and Yuxin Wen. Why does Feedback-Augmented Self-Distillation fail to improve Retrieval-Interleaved search agents? *arXiv preprint arXiv:2607.17558*, 2026. URL https://arxiv.org/abs/2607.17558.

Tianzhu Ye, Li Dong, Guanheng Chen, He Zhu, Xun Wu, Shaohan Huang, and Furu Wei. LLM-as-a-Coach: Experiential learning for Non-Verifiable tasks. *arXiv preprint arXiv:2607.18110*, 2026a. URL https://arxiv.org/abs/2607.18110.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *ArXiv*, abs/2602.12275, 2026b.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025. URL https://arxiv.org/abs/2508.03905.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2401.10020.

Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. Usimagent: Large language models for simulating search users. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. URL https://arxiv.org/abs/2403.09142.

Junjie Zhang, Guozheng Ma, Shunyu Liu, Zetian Hu, Yongcheng Jing, Ting-En Lin, Yongbin Li, and Dacheng Tao. STRIDE: Learnable stepwise language feedback for LLM reasoning. *arXiv preprint arXiv:2605.18851*, 2026. URL https://arxiv.org/abs/2605.18851.

Xiaoying Zhang, Yipeng Zhang, Hao Sun, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *arXiv preprint arXiv:2506.03106*, 2025a. URL https://arxiv.org/abs/2506.03106.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. BREAD: Branched rollouts from expert anchors bridge SFT & RL for reasoning. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2506.17211.

Binbin Zheng, Zijun Xie, Guanqun Zhao, Enlei Gong, Xing Ma, Xiaoliang Fu, and Zeyu Chen. Group-Reflective Self-Distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.28076*, 2026. URL https://arxiv.org/abs/2607.28076.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank F. Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, Ronan Le Bras, and Maarten Sap. HAICOSYSTEM: An ecosystem for sandboxing safety risks in human-ai interactions. In *Second Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2409.16427.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. Social world models, 2026a. URL https://arxiv.org/abs/2509.00559.

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.

A Appendix

Limitations. Our approach has several limitations. First, both the training feedback and part of the evaluation rely on LLM judges as proxies for human judgment, so improvements may partly reflect judge preferences. Second, DITTO doubles the rollout cost per prompt (a teacher rollout for each student rollout) and adds one judge call per rollout. Third, our experiments use a single base model family (Qwen3) and English-only benchmarks.

B Training Configurations and Robustness

Table 2 lists every training configuration we evaluated on Sotopia. We first selected a GRPO recipe on Sotopia and ToM, then applied it unchanged to GRPO, DITTO, and all ablations; DITTO received no method-specific tuning. The same table reports DITTO across four random seeds (4.69 ± 0.02) and on a smaller Qwen3-4B base (4.51, still above 8B GRPO at 4.23).

Hyperparameters. LoRA rank 32 and alpha 64; constant learning rate 5e-6; batch size 64 with mini-batch size 16 (4 optimization steps per batch); group size 8; maximum prompt and output lengths 8192; about 100 training steps; DAPO-style token-level loss with a decoupled (clip-higher) ratio clipping range and no KL penalty. For multi-agent tasks (CoSER), the final reward is assigned to each agent. In preliminary experiments, LoRA and full-parameter training performed comparably, so we use LoRA for efficiency.

B.1 Human annotation

We collected annotations from 76 distinct online annotators on Prolific. Participants ranged in age from 21 to 78 years old (median = 39; mean 40.8 ± 12.8). The sample included 63% female and 35% male annotators, with the remainder not disclosing this information. In terms of ethnicity, the

Table 2: Training configurations evaluated on Sotopia (validation eval_avg at step 120; higher is better). The GRPO recipe was selected on Sotopia and ToM from candidates (a)–(i); the final setup (g) is then used unchanged for GRPO, DITTO, and all ablations on all tasks. (h) scores higher but at roughly 2× the compute of (g), so we keep (g) as the shared configuration. †Fixes to the Sotopia environment harness (episode length, environment prompts, and turn-handling logic), applied to all subsequent rows.

Group	Configuration	Score
Prior work	Sotopia-RL-7B (GRPO + process reward) (Yu et al., 2025)	3.12
GRPO recipe	(a) Seed-OSS-36B + PPO	2.62
	(b) Seed-OSS-36B + GRPO	3.25
	(c) (b) + KL penalty	2.86
	(d) (b) + batch size 16	2.96
	(e) Qwen3-8B + GRPO + LoRA (rank 16) + env fixes†	3.71
	(f) (e) + clip-higher + token-mean loss	3.90
	(g) (f) + no KL + LoRA (rank 32) *final setup*	4.23
	(h) (g) + batch size 128	4.38
	(i) rejection fine-tuning on (h)	4.17
DITTO	DITTO, config (g), seeds 1 / 2 / 3 / 4	4.71 / 4.69 / 4.72 / 4.65
	mean ± std over seeds	4.69 ± 0.02
	DITTO, Qwen3-4B base	4.51
Loss ablations	DITTO only (y_1, x)	2.57
	DITTO no $\mathcal{L}_{\text{fb}}$	3.46
	RLTF-SD (Song et al., 2026)	3.59
	ERL (Shi et al., 2026a)	3.59
Reverse distillation	SDPO+ token	2.94
	SDPO+ logits	2.90
	SDPO token	0.97
	SDPO logits	0.22

pool consisted of White (45%), Black (24%), Asian (17%), Mixed (4%), and other (4%) annotators, with the remainder undisclosed. All annotators resided in the United States, and the majority reported English as their primary language (85%). Most participants reported not being current students (71%, vs. 12% students; the rest undisclosed).

Each annotator was assigned one task group: the three model-simulated versions of the same task, presented in random order. Each conversation opens in a web-based chat interface (Figure 6). The right panel shows the annotation instructions and the task context. For Sotopia, this includes the scenario, the characters' relationship, etc. The left panel shows the completed conversation and the survey. For Sotopia, each rating dimension is a slider with its own definition and score range, as in the original benchmark rubric. Each benchmark presents a different annotation survey according to its task rubric.

C TRAINING CURVE

See Figure 9, Figure 10, and Figure 11 for training curves. See Table 4 for the complete baseline results. Figure 7 reports the human evaluation of Section 6.3 with all per-dimension scores.

Figure 8 shows the teacher–student score gap on all Sotopia dimensions.

Table 3 gives the full verbal feedback condensed in Figure 5.

D SOUL TASK DETAILS

This appendix provides full descriptions, evaluation protocols, and examples for each of the 10 SOUL tasks. Table 5 summarizes all tasks, and Table 7 details the training data sources. Table 6 surveys human-validation studies for benchmarks using LLM-as-judge evaluation.

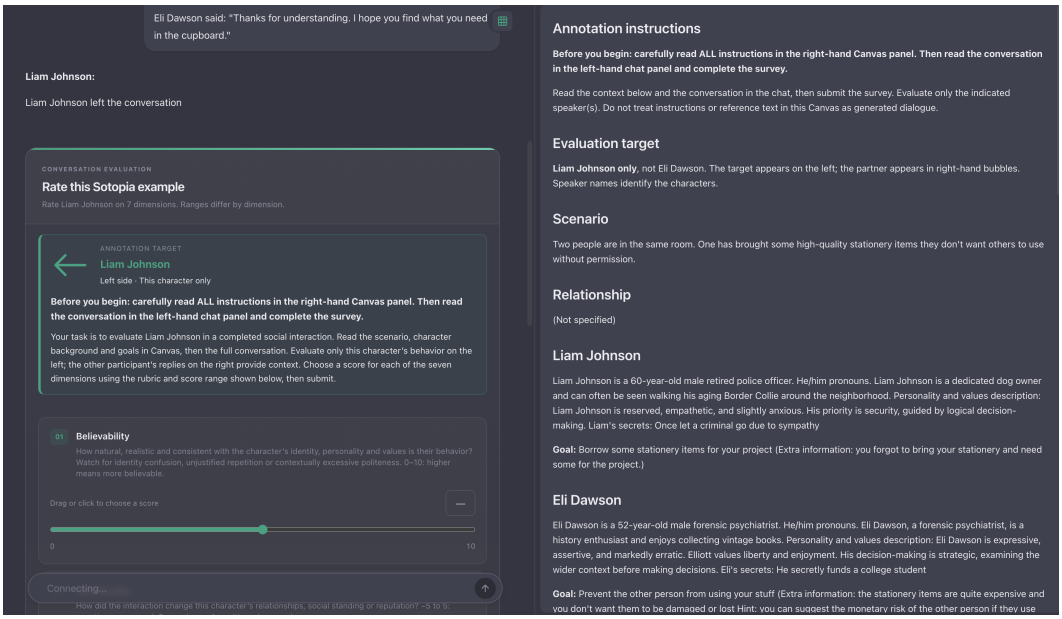


Figure 6: A Sotopia annotation interface example. The left panel shows the simulated conversation and the survey that follows it. For Sotopia, the annotator rates only the target character on seven Sotopia dimensions, each with its own score range slider (e.g., believability 0–10).

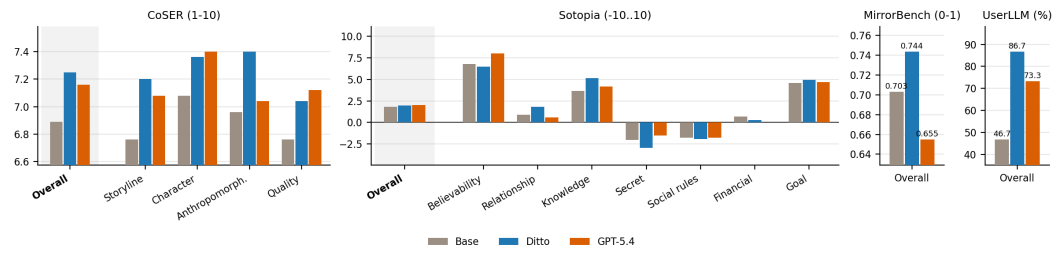


Figure 7: Human evaluation on the four judge-scored SOUL tasks, with per-dimension scores for CoSER and Sotopia. Each task keeps its own scale; the shaded group is the overall score.

Train/test separation. No training instance overlaps with any test instance. For tasks adopted from existing benchmarks we follow their official splits. For tasks built from related sources we separate at the source level: LifeChoices and CoSER are both derived from novels, but their test splits use disjoint novels, and we removed from training any source material (e.g., held-out novels) that overlaps with a test source.

D.1 THEORY OF MIND

FanToM (Kim et al., 2023) stress-tests Theory of Mind in information-asymmetric multi-party conversations. Characters join and leave conversations at different points, creating natural knowledge gaps. The benchmark poses 1,086 questions across three categories—belief, answerability, and information accessibility—in both multiple-choice and free-response formats.

HiToM (He et al., 2024) evaluates higher-order Theory of Mind reasoning, requiring models to reason about nested beliefs (e.g., "Alice thinks that Bob thinks the object is in location X"). Scenarios include deception and varying communication patterns (private vs. public).

ToMi (Nematzadeh et al., 2018) is one of the most classic tests of Theory of Mind. We use the more challenging version by Sclar et al. (2023), which extends the paradigm through diverse paraphrasing

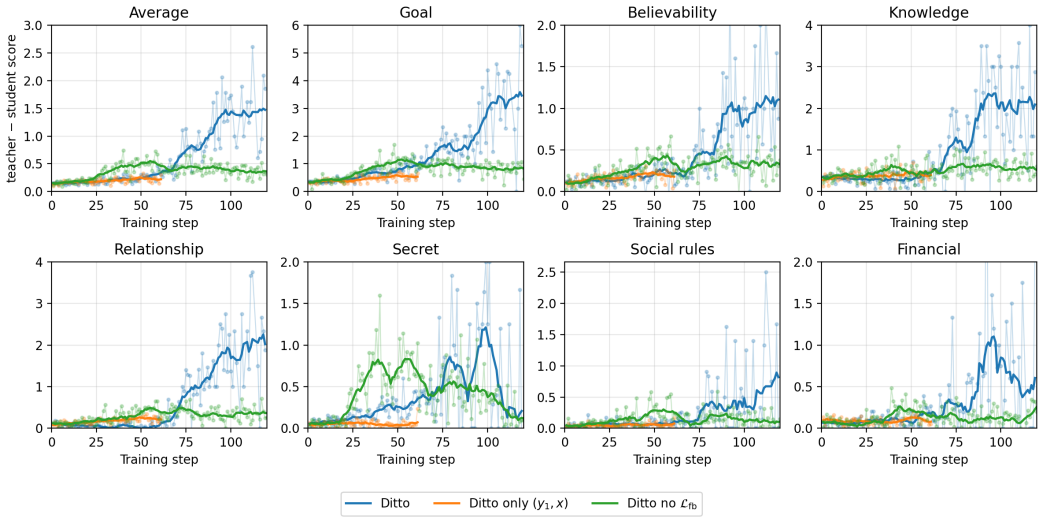


Figure 8: Teacher–student score delta on all Sotopia dimensions during training.

	Step 10		*Step 100*	
Dimension	**Score**	**Feedback**	**Score**	**Feedback**
Goal	**1**	The score is low because the session ended without any purchase and **you never created a path that could realistically lead Samuel to accept your floor**. The key omission holding you back is that you never used structuring tactics that could convert "I can't do $17.3 today" into "OK—can we do X that still lands on $17.3?"	**3**	The score was held back because **no concrete agreement meeting $17.30 happened**, even though you came close by proposing $15 + bonus math. The key omission was: you never found a structure that both (a) respects the **"cash-only, no strings" constraint** Samuel demands and (b) guarantees your seller minimum $17.30 in the simplest terms he can accept in the moment.
Relationship	**-1**	The score was negative because the exchange stayed **purely adversarial/rigid with repeated dismissal**. Samuel tried to be respectful and clear ("I hear you… I respect the stance"), but you responded with more cutting language ("Don't waste my time again," "I'm not here to entertain second chances").	**1**	Relationship scoring here rewards **warmth, trust, and clarity**—showing appreciation without escalating into a long, "theatrical" negotiation spiral. A +4/+5 outcome would include: you acknowledge Samuel's integrity, **reduce friction by making the deal structure simple**, and show willingness to walk away respectfully without pressure.
Knowledge	**2**	You did learn Samuel's budget cap (he offered $14 and $16) and that he can't meet $17.3 today, but you missed other potentially useful info: **timing, ability to withdraw cash, whether he can arrange a partial earlier cash + remainder later while still totaling $17.3**, or whether he's willing to buy if you bundle pickup logistics.	**6**	You got some of this, but you didn't lock down one critical point: **whether his refusal is about price total, about structure, or about trust/complexity**. Without that, you keep proposing structures that can't pass his real constraint.
Believability	**7**	Believability was solid because you were direct and consistent, but it dipped because of **over-repetition and occasionally overly theatrical or lecture-like phrasing** ("small miracle," "lesson in markets," "you've made my position clear… done talking to you" repeatedly).	**7**	The judge noted you were plausible overall, but penalized for **stylized, escalating bonus schemes and repeated "theatrical" logic**. Believability improves when you sound like you're running a transaction like a chef runs service: **clear, direct, confident, not overly convoluted**.
Financial	**0**	Here, **no sale occurred**, so material outcome was neutral in the rubric. You missed an opportunity to convert "close but short" into a realistic path that lands at $17.3.	**-2**	Negative means you likely lost value by failing to close. You tried to engineer ways to reach $17.30, but Samuel refused any **bonus/credit structures** and also refused your **"blended" cash notion**.

Table 3: Full verbal feedback on the same Sotopia scenario at training steps 10 and 100 (condensed in Figure 5).

of ToMi narratives. By varying surface forms of identical scenarios, it tests whether models exhibit robust Theory of Mind reasoning rather than exploiting shallow textual cues. We evaluate on 600 questions covering reality, memory, and first/second-order belief.

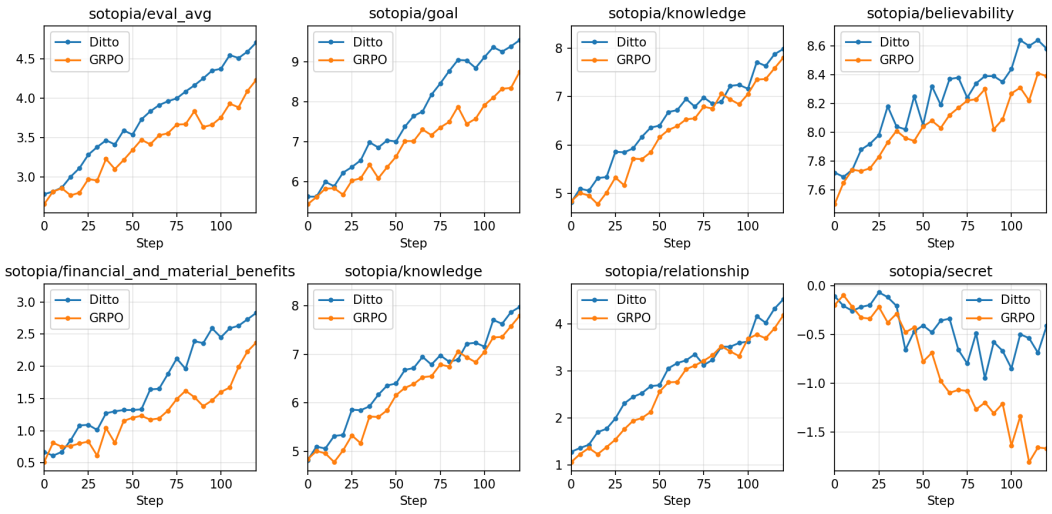


Figure 9: Training results on Sotopia

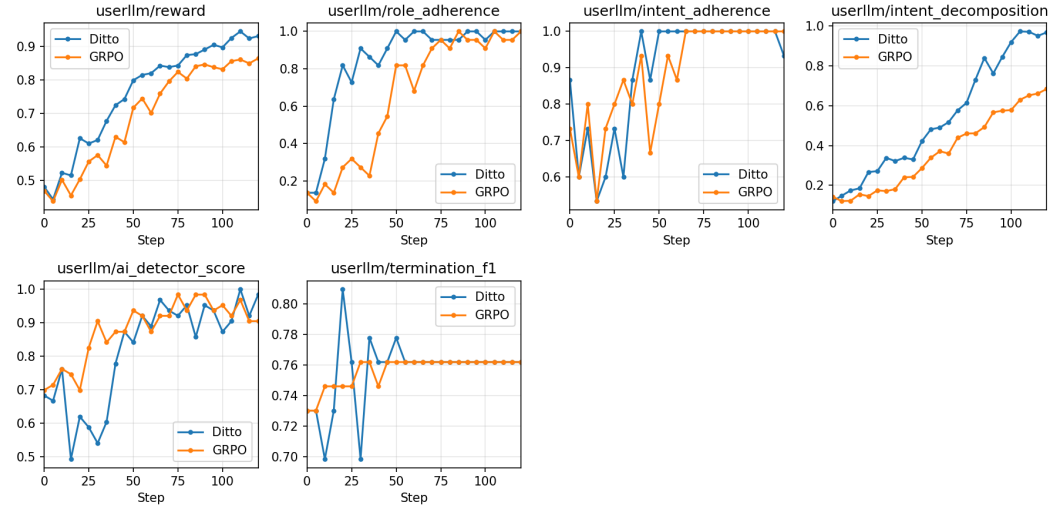


Figure 10: Training results on UserLLM

D.2 Character Role Play

CoSER (Wang et al., 2026b) is a literary character role-playing benchmark built from 771 books containing 17,966 characters. Using a "given-circumstance acting" methodology inspired by theater, models must simulate multi-turn dialogues as literary characters in specified narrative settings. Evaluation uses LLM judges that identify specific flaws across four dimensions—Storyline Consistency, Anthropomorphism, Character Fidelity, and Storyline Quality—with severity ratings from 1 to 5. We evaluate on both in-distribution (ID, 100 conversations from seen books) and out-of-distribution (OOD, 100 conversations from unseen books) splits (see Section D.7 for the full generation protocol, scoring formula, and examples).

LifeChoices (Xu et al., 2024a) evaluates persona-driven decision making for literary characters. Given a character biography, scenario, and candidate options, the model must predict the decision that best matches the character's original choice in the narrative. The dataset contains 1,462 decision points from 388 novels; we evaluate on 600 tasks using accuracy (see Section D.8 for task format, context length statistics, and examples).

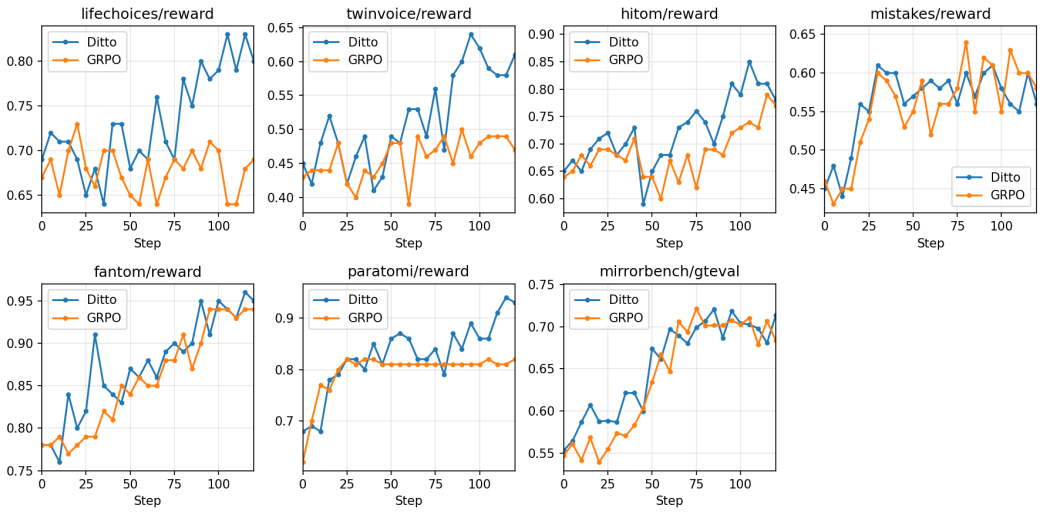


Figure 11: Training results on Soul tasks

Benchmark	GPT-5.4	GPT-5 nano	Sotopia-RL 7B	HumanLM 8B	HER 32B	Coser 8B	Base 8B	GRPO 8B	DITTO 8B
FanToM	0.900	0.720	0.150	0.780	0.550	0.030	0.780	0.940	**0.950**
HiToM	0.700	0.370	0.320	0.510	0.560	0.020	0.580	0.770	**0.780**
ToMi	0.880	0.850	0.610	0.720	0.630	0.430	0.680	0.820	**0.930**
CoSER	**0.659**	0.352	0.296	0.227	0.419	0.305	0.435	0.541	0.512
LifeChoices	**0.870**	0.600	0.670	0.660	0.750	0.520	0.670	0.690	0.800
Sotopia	0.300	0.310	0.312	0.275	0.290	0.241	0.277	0.423	**0.470**
Mistakes	0.570	**0.580**	0.180	0.520	0.350	0.000	0.460	**0.580**	0.560
MirrorBench	0.536	0.358	0.367	0.481	0.361	0.314	0.547	0.683	**0.713**
UserLLM	0.575	0.324	0.404	0.318	0.537	0.420	0.469	0.863	**0.930**
TwinVoice	**0.640**	0.230	0.340	0.380	0.390	0.070	0.430	0.470	0.610
Average	0.663	0.469	0.365	0.487	0.484	0.235	0.533	0.678	**0.726**

Table 4: **Main results on SOUL tasks.** We report the primary metric for each benchmark (higher is better). For Sotopia, we report the normalized eval_avg score. *Base* denotes the base model, Qwen3-VL-8B-Instruct. **Bold** indicates the best result in each row.

D.3 SOCIAL SKILL

Sotopia (Zhou et al., 2024) evaluates social interaction through two-agent simulations in realistic scenarios. Each agent receives a character background, social goals, and relationship context. Performance is measured across seven dimensions—believability, relationship, knowledge, secret-keeping, social rules, financial outcomes, and goal achievement—scored by LLM judges.

D.4 LEARNER SIMULATION

Mistakes (Ross & Andreas, 2025) evaluates whether LLMs can simulate *incorrect* student reasoning in K-12 mathematics. Given a multiple-choice math problem and a misconception description, models must select the wrong answer that corresponds to that specific misconception. This tests a key aspect of user simulation: faithfully reproducing systematic human errors.

D.5 USER SIMULATION

MirrorBench (Hathidara et al., 2026) evaluates user-proxy agents on producing human-like utterances, decoupled from downstream task success. Given an extracted user goal, the model iteratively generates all user turns in a conversation while assistant responses are replayed from the original

Table 5: SOUL evaluation tasks. Each row lists the task, a brief description, evaluation format, number of test instances, and primary metric. We evaluate on a fixed random subset of up to 100 instances per task (Section 5).

Benchmark	Task	Format	# Test	Metric
Theory of Mind				
FanToM	ToM in multi-party conversations	MCQ + QA	1,086	Accuracy
HiToM	Higher-order nested belief tracking	MCQ	100	Accuracy
ToMi	False-belief QA with paraphrased stories	QA	600	Accuracy
Character Role Play				
CoSER (ID/OOD)	Literary character dialogue	Multi-turn gen.	200	LLM judge
LifeChoices	Persona-driven character decisions	MCQ (4-way)	600	Accuracy
Social Skill				
Sotopia	Social interaction with goals	Multi-agent int.	100	LLM judge
Learner Simulation				
Mistakes	Select answer matching a misconception	MCQ (4-way)	600	Accuracy
User Simulation				
MirrorBench	Generate human-like user turns	Multi-turn gen	795	LLM judge
UserLLM	Generate user turns (multi-metric)	Single-turn gen	1,800	Multi-metric
Persona Simulation				
TwinVoice	Select answer matching the user's persona	MCQ (4-way)	5,687	Accuracy

Table 6: Human evaluation validation for benchmarks using LLM-as-judge. We report whether each benchmark conducted a human study to validate its LLM judge, the study design, and the reported agreement. Benchmarks using only automatic metrics (accuracy, substring match, etc.) are omitted.

Benchmark	Human Study Design	Agreement	Validated?
CoSER	60 samples, 3 annotators, 1–10 scale	68.6% (GPT-4o), 77.5% (DeepSeek-R1) alignment with human judges	✓
Sotopia	Pearson correlation with human labels across 7 dimensions	GPT-4 validated as proxy; strongest on goal completion, finances, relationships	✓
MirrorBench	100 episodes per metric, stratified, blinded annotators	GTEval: strong; Pairwise Indistinguishability: moderate (Spearman ρ, Pearson r, Kendall τ; all $p \ll 0.001$)	✓
UserLLM (NQ)	None reported	GPT-4o used as judge; no human correlation or agreement metrics reported	×

data. Each generated turn is compared against the real human utterance. Evaluation uses both lexical diversity metrics (MATTR, Yule's K, HD-D) and LLM-judge metrics, calibrated against human-human and proxy-proxy baselines. We evaluate on four subsets: ChatbotArena, ClariQ, OASST1, and QuLAC (see Section D.9 for detailed dataset statistics and examples).

UserLLM (Naous et al., 2025) evaluates single-turn user utterance generation. Given a conversation context and a user intent, the model generates one user turn. We evaluate on three test sets, each targeting different metrics (see Section D.10 for detailed metric definitions and examples):

- **CSQA (Role Adherence):** The model plays a user who asks a commonsense question. Role adherence is measured by a *substring-matching* rule: the answer-choice texts are normalized and checked against the generated turn. If the turn contains one or two (but not all) of the answer choices, it scores 0—the model is leaking specific answers rather than staying in the questioner role. If it mentions none, it scores 1.
- **NQ (Intent Adherence):** The model plays a user asking a factoid question. Intent adherence is measured by an *LLM judge* that classifies whether the generated turn refuses to answer the question (as a real user would) or inappropriately provides the answer itself.

Table 7: SOUL training data sources.

Category	Benchmark	Data Source	Format	Size
Theory of Mind	FanToM	Train split	MCQ context	–
	HiToM	Augmented	MCQ context	–
	ToMi	Train split	QA context	–
Character Role Play	CoSER	Train books	Multi-turn	–
	LifeChoices	Train split	MCQ context	–
Social Skill	Sotopia	Scenario pool	Multi-agent	–
Learner Sim.	Mistakes	Train split	MCQ context	–
User Sim.	MirrorBench	Held-out episodes	Multi-turn	–
	UserLLM	Train split	Single-turn	–
Persona Sim.	TwinVoice	Bluesky user posts	Persona histories + MCQ (4-way)	471
		Pushshift Telegram messages	Persona histories + MCQ (4-way)	384
		Project Gutenberg novels	Persona histories + MCQ (4-way)	480

- **PRISM (Multi-metric):** The model generates first turns from high-level intents. Four metrics are computed: *first-turn diversity* (pairwise 1-gram Jaccard distance across generations), *intent decomposition* (1-gram overlap between intent keywords and the generated turn), *termination F1* (whether the model correctly predicts conversation ending), and *AI detector score* (whether the text passes an AI-detection classifier as human-written).

D.6 PERSONA SIMULATION

TwinVoice (Du et al., 2025) is a multi-dimensional benchmark for persona simulation spanning Social Persona, Interpersonal Persona, and Narrative Persona. Given a persona history and a new stimulus problem, the model must identify the response that best matches the target persona in a discriminative 4-way multiple-choice task. The benchmark contains 5,687 test instances in total and reports accuracy, while also organizing examples around six fine-grained capabilities, including opinion consistency, memory recall, logical reasoning, lexical fidelity, persona tone, and syntactic style.

D.7 COSER

CoSER (Wang et al., 2026b) is a literary character role-playing benchmark built from 771 books containing 17,966 characters. Using a "given-circumstance acting" (GCA) methodology inspired by theater, models must simulate multi-turn dialogues as literary characters in specified narrative settings. We describe the generation protocol, evaluation metrics, and dataset details below.

Generation protocol. Each task (called a *circumstance*) specifies a scene from a book, including: (1) a **scenario** describing the setting, atmosphere, and dramatic tension; (2) a **plot summary** providing narrative context; (3) detailed **character profiles** for each speaking character; and (4) **character motivations** (inner thoughts and goals). The model must simulate a multi-character conversation in this setting.

The **actor model under evaluation** plays *all* characters in the conversation—the same LLM sequentially portrays each character $c \in \mathcal{C}$ in the given setting. Two auxiliary agents support the simulation:

- **Character agents** (actor model): One instantiation per speaking character. Each receives the scenario, its own profile and motivation, and the profiles of other characters. It generates dialogue in the format `[inner_thought] speech (action)`.
- **Environment agent** (auxiliary): Describes environmental changes and minor character reactions based on the ongoing dialogue (e.g., crowd murmurs, weather shifts).
- **Next-speaker predictor (NSP)** (auxiliary): After each turn, predicts which character speaks next (or outputs `<END CHAT>` to terminate). Termination requires a minimum of 5 rounds.

Conversations run for up to 20 rounds. In practice, the NSP terminates most conversations earlier (mean: 14.2 character turns, range: 6–20).

The generated conversation is then compared against the **ground-truth dialogue** from the original book, which serves as the reference for evaluation.

Evaluation protocol. Evaluation combines LLM-based flaw detection with automatic lexical metrics. For each of four dimensions, an LLM judge (acting as a "literary critic") receives the full context—book title, plot summary, scenario, character profiles, the reference conversation, and the dimension-specific evaluation criteria—and identifies specific *flaws* in the simulated conversation. Each flaw is assigned a severity from 1 (minor) to 5 (severe).

The per-dimension score is computed as:

$$\text{score} = \max\Big(0,\ \min\big(100 - (\text{total_severity} - 0.3 \times N_{\text{turns}}) \times 5,\ 100\big)\Big)$$

where N_{turns} is the number of character turns in the simulation. The $0.3 \times N_{\text{turns}}$ term provides a length-based tolerance: longer conversations are expected to accumulate more minor flaws. The overall reward is the mean of four dimension scores, normalized to $[0, 1]$.

Additionally, **BLEU** and **ROUGE-L** are computed against the reference dialogue, though these are typically low (BLEU ≈ 0.02) since valid simulations can differ substantially in wording from the original.

Evaluation dimensions. The four dimensions capture different facets of role-playing quality:

- **Storyline Consistency**: Whether characters' reactions (emotions, attitudes, behaviors) align with those in the reference conversation. Detects deviations from the original dramatic arc.
- **Anthropomorphism**: How human-like and natural characters behave. Checks for: self-identity (goals, decisions, preferences vs. "helpful AI assistant" behavior), emotional depth (psychological complexity, use of subtext), persona coherence (consistent personality), and social interaction (understanding of others, contextual reactions).
- **Character Fidelity**: How well characters match their established profiles from the book. Evaluated only for major characters. Checks: appropriate language/vocabulary for the character's background, character-specific knowledge, personality-consistent emotions and behaviors, and appropriate social dynamics given relationships and status.
- **Storyline Quality**: Logical consistency and narrative quality. Checks: natural flow and progression (no verbosity, redundancy, or repetition), and factual consistency between statements.

Human validation. The original paper validates the LLM-based evaluation against human judgments. Three annotators scored 7 representative LLMs on 60 samples using a 1–10 scale. Human rankings generally align with LLM-judged scores: CoSER-70B ranks highest under both human (6.78 avg, 86.9% win rate) and LLM evaluation. The full GCA evaluation protocol achieves 68.6% alignment with human judges (GPT-4o as critic); DeepSeek-R1 as critic achieves 77.5%. Ablation analysis shows all GCA components contribute: removing reference dialogues, scoring rubrics, length correction, or dimension separation each reduces alignment. Interestingly, simple n-gram metrics (BLEU: 75.3%, ROUGE-L: 72.0%) achieve competitive alignment with LLM judges, though they cannot capture the nuanced flaw-level feedback that GCA provides.

Datasets. We evaluate on two splits. Table 8 summarizes their structure.

Table 8: CoSER dataset statistics.

Split	# Tasks	Avg GT turns	Description
ID (in-distribution)	100	14.2	Scenarios from books seen during training
OOD (out-of-distribution)	100	14.2	Scenarios from unseen books

Example. Figure 12 shows a representative CoSER task from *Pride and Prejudice*.

Table 9: CoSER input context lengths (characters). "Scenario" includes the scene description; "Profiles" is all character profiles combined; "Total" is scenario + profiles + plot summary + character motivations.

Split	Avg total	Median	P90	Max
ID	7,419	7,092	10,170	13,738
OOD	7,458	6,746	10,916	18,146

CoSER — Example (Pride and Prejudice)

Scenario: *The scene is set in Mr. Bennet's private study. Elizabeth has been summoned unexpectedly. Mr. Bennet holds a letter from Mr. Collins about Elizabeth's rumored engagement to Mr. Darcy.*

Characters: Mr Bennet (sardonic patriarch), Elizabeth Bennet (spirited, anxious about rumors)

Ground-Truth Dialogue

Mr Bennet: [I can barely contain my amusement.] Lizzy, I was going to look for you; come into my room.

Elizabeth: [I feel a sense of dread but must appear calm.] (follows her father) What is it, Papa?

Mr Bennet: [This will be entertaining.] I have received a letter this morning that has astonished me exceedingly…

Model-Generated Dialogue (GPT-4o)

Mr Bennet: (looks up from the letter with a wry smile) Ah, Elizabeth, my dear. It seems we have a rather amusing piece of correspondence here.

Elizabeth: (leans forward with curiosity) Father, I can only imagine the vigor with which she has expressed her disapproval…

Mr Bennet: (hands the letter to Elizabeth with a chuckle) She insists that you must not marry Mr. Darcy…

Evaluation Scores (0–100):

Storyline Consistency	69.5	Anthropomorphism	69.5
Character Fidelity	34.5	Storyline Quality	59.5

Overall Reward: 0.583 BLEU: ≈0 ROUGE-L: 0.237

Key flaws identified: Mr Bennet is portrayed as warm and supportive rather than characteristically sardonic and detached (Character Fidelity: severity 4). Emotional tension is reduced compared to the original (Storyline Consistency: severity 3). The conversation lacks Mr Bennet's signature sarcasm (Anthropomorphism: severity 4).

Figure 12: CoSER example from *Pride and Prejudice*. The model captures the general narrative arc (reading Mr. Collins' letter) but softens Mr Bennet's distinctive sardonic personality into generic warmth, resulting in low Character Fidelity (34.5). The judge identifies this as the primary failure mode: the character behaves too much like a "helpful AI assistant" rather than the complex, detached patriarch of the original.

D.8 LIFECHOICES

LifeChoice (Xu et al., 2024b) tests persona-driven decision-making by presenting literary characters with pivotal life decisions from 388 novels. Models must predict what a character *would* do based on their established personality, narrative context, and motivations. We describe the task format, evaluation, and dataset details below.

Task format. Each task provides: (1) a **character biography** (`input_text`)—an extensive narrative summary of the character's background, personality, experiences, and arc extracted from the source novel; (2) a **scenario** describing the specific situation the character faces; (3) a **question** asking what decision the character makes; and (4) **four options** (A–D), one of which matches the character's actual choice in the novel. Each option is accompanied by a **motivation** explaining the psychological reasoning behind that choice.

The model receives the character biography, scenario, question, and options as a structured prompt, and must select the correct answer letter.

Evaluation. Evaluation is straightforward **exact-match accuracy**: the model's selected option is compared to the ground-truth answer. No partial credit is given. The original paper reports human performance of 92% versus 63% for the best LLM baseline (GPT-4), indicating a substantial gap in persona-driven reasoning.

Context lengths. LifeChoice tasks have notably long input contexts due to the extensive character biographies. Table 10 summarizes the context length distribution.

Table 10: LifeChoice input context lengths (characters). The context consists of the character biography, scenario, question, and options concatenated.

	Avg	Median	P90	Max
Total context	26,177	25,513	46,300	71,773

The average context of ∼26K characters (∼6.5K tokens) is among the longest in our benchmark suite, reflecting the rich narrative information needed for persona-driven reasoning.

Dataset. The full dataset contains 1,462 decision points from 388 novels. We evaluate on a sample of 600 tasks. Decision points span diverse literary genres (fantasy, mystery, literary fiction, science fiction) and character archetypes.

Overlap with CoSER. Both LifeChoice and CoSER (Appendix D.7) draw from literary novels but test different capabilities: LifeChoice is a discriminative MCQ task (predicting what a character *would* do), while CoSER is a generative multi-turn dialogue task (acting *as* a character). We find 39 books (out of LifeChoice's 383) that also appear in CoSER's full 772-book corpus, covering 151 LifeChoice tasks (9.6%). Of these, only 3 books overlap with CoSER's *test* set (16 tasks, 1.0%); the remaining 36 books overlap with CoSER's training-only partition (135 tasks, 8.6%). This overlap is a potential data contamination concern for models fine-tuned on CoSER training data but does not affect our evaluation of general-purpose LLMs.

Example. Figure 13 shows a representative LifeChoice task.

D.9 MIRRORBENCH

MirrorBench (Hathidara et al., 2026) evaluates user-proxy agents on producing human-like utterances, decoupled from downstream task success. We describe the generation protocol, evaluation metrics, our adaptation, and dataset details below.

Generation protocol. Given a reference dialogue $d_j^{\text{ref}} = [(u_{j,1}^{\text{ref}}, a_{j,1}^{\text{ref}}), \ldots, (u_{j,L}^{\text{ref}}, a_{j,L}^{\text{ref}})]$ and a synthesized user goal g_j summarizing the user's intent and communication style, MirrorBench generates a synthetic rollout $\hat{d}_j = [(\hat{u}_{j,1}, \hat{a}_{j,1}), \ldots, (\hat{u}_{j,L}, \hat{a}_{j,L})]$ as follows. At each turn t, the user-proxy LLM θ_u receives the accumulated dialogue history $[(\hat{u}_{j,1}, \hat{a}_{j,1}), \ldots, (\hat{u}_{j,t-1}, \hat{a}_{j,t-1})]$ plus the goal g_j as a system prompt, and generates the next user turn $\hat{u}_{j,t}$. The assistant LLM θ_a then produces a contextual reply $\hat{a}_{j,t}$, conditioned on both the ongoing synthetic dialogue and the original reference conversation d_j^{ref}, so that the rollout trajectory stays close to the original. Critically, *all* user turns are generated by the proxy and *all* assistant turns are generated by an LLM—there is no verbatim replay of original turns. In the original paper, the assistant is fixed to GPT-4o for all primary experiments; the authors report that "LLM hallucinations can occasionally induce drift" in assistant responses but find divergence in $<1\%$ of samples.

Evaluation granularity. All metrics operate at the **full-conversation level**, not turn-by-turn. The judge sees the entire proxy conversation $\hat{d}_j$ (or its user-side transcript) compared against the entire reference d_j^{ref}, and produces a single score per episode. For lexical metrics, all proxy user turns are concatenated into one token sequence $\hat{\mathbf{t}}_i$ and compared against the concatenated human user turns $\mathbf{t}_i^{\text{ref}}$.

LifeChoice — Example

Character: Amelia Choquet **Book:** *A Great Reckoning*

Character Biography (excerpt, 300 of ∼25K chars):

The novel begins in November in the small Canadian town of Three Pines. Armand Gamache is at home, reviewing applications from individuals hoping to be admitted as students into the police academy. The file of one young woman named Amelia Choquet sticks in his mind. . .

Scenario: *Amelia Choquet, a cadet at the Sûreté Academy, is faced with a proposition from her professor, Serge Leduc, who offers to secure her position at the academy in exchange for sexual favors. Upon refusal, Leduc threatens to expel her.*

Question: *What decision does Amelia make when confronted with Leduc's proposition and threats?*

Options:

A. Accept Leduc's proposition to secure her future at the academy.
B. Reject Leduc's advances and refuse to be intimidated by his threats.
C. Report Leduc's behavior to the academy's administration, risking disbelief and retaliation.
D. Attempt to transfer to another academy to escape the situation.

Correct Answer: B

Motivation for B: *Amelia resolutely rejects Leduc's advances because her integrity and self-respect are paramount. She refuses to be coerced, despite the threat of expulsion. . .*

Figure 13: LifeChoice example. The model receives ∼25K characters of character biography and must predict Amelia's decision. All four options are psychologically plausible—the distractors represent reasonable alternatives—but only one matches the character's actual choice in the novel. Selecting the correct answer requires understanding Amelia's personality, values, and character arc.

Metrics. MirrorBench combines two metric families: three *lexical diversity* metrics and three *judge-based realism* metrics, capturing complementary facets of human-likeness.

LEXICAL DIVERSITY METRICS measure vocabulary richness and repetition patterns of the proxy's user turns, z-score normalized against the human distribution on the same dataset:

- **MATTR** (Moving-Average Type-Token Ratio): vocabulary diversity over a sliding window ($w = 50$).
- **HD-D** (Hypergeometric Distribution Diversity): vocabulary diversity robust to sequence length.
- **Yule's K**: repetitiveness from the token-frequency spectrum (higher K = more repetition).

JUDGE-BASED REALISM METRICS use LLM judges to assess whether the proxy's user turns *sound* like a real human. The three metrics differ in what information the judge receives and what question it answers:

- **GTEval** (reference-comparative): The judge sees *both* the proxy conversation and the human reference side by side, and scores their similarity on $[0, 1]$. This measures how closely the proxy reproduces the specific human behavior.
- **Pairwise Indistinguishability (PI)** (reference-discriminative): The judge sees two anonymized conversations—one proxy, one human—in random order (labeled A/B) and picks which user sounds more human. This is a blind A/B test; the score is the proxy's win rate. A score near 0.5 means the proxy is indistinguishable from the human.
- **Rubric-and-Reason (RNR)** (reference-free): The judge sees *only* the proxy conversation and rates whether the user sounds human-like based on a rubric covering style, behavior, and tone. Returns YES/NO. This judges realism in absolute terms without any human reference.

Each judge metric is executed with a self-consistency parameter $c \geq 1$ (repeated judgments under the same prompt). Scores are calibrated via Human–Human (HH) and Proxy–Proxy (PP) controls: HH

compares the human conversation to itself (expected ceiling ≈ 1.0 for GTEval, ≈ 0.5 for PI), and PP compares the proxy to itself (expected baseline).

Our adaptation. In our evaluation framework, we run MirrorBench in *single-turn assistant mode*: the user-proxy LLM generates all user turns iteratively, but assistant responses are replayed verbatim from the original dataset rather than generated by a separate LLM. This simplification is inconsequential for conversations with a single user turn (most of Arena, all of QuLAC), since there is no subsequent turn where assistant coherence matters. For multi-turn conversations (all of ClariQ, most of OASST1), replayed assistant responses may become contextually misaligned if the proxy's user turn diverges from the original, potentially causing cascading drift. We use GPT-5.4 as the judge model for all metrics.

Datasets. We evaluate on four subsets drawn from different conversational domains. Table 11 summarizes their structure.

Table 11: MirrorBench dataset statistics. "User turns" indicates how many user messages the model must generate per conversation.

Subset	# Tasks	User turns	Total turns	Domain
Arena	195	1–12 (avg 1.2)	2–24	Open-ended chat
ClariQ	200	4 (all)	7	Query clarification
OASST1	200	1–3 (avg 1.8)	2–6	Open-domain QA
QuLAC	200	1 (all)	2	Query clarification

Table 12: MirrorBench input context lengths (characters). "Task desc." is the synthesized user goal; "Conv." is the reference conversation content. Percentiles are over total context (task description + conversation).

Subset	Avg desc.	Avg conv.	Avg total	P50	P90	Max
Arena	307	1,172	1,479	1,179	2,615	14,976
ClariQ	336	327	663	658	755	848
OASST1	339	1,811	2,150	1,862	3,931	11,648
QuLAC	302	93	394	388	468	546

- **ChatBot Arena** (Zheng et al., 2023) conversations come from the LMSYS Chatbot Arena platform. Most (173/195) are single user-turn exchanges, but 22 are multi-turn (up to 12 user turns).
- **ClariQ** (Aliannejadi et al., 2019) consists of information-seeking dialogues where the user progressively clarifies their intent across 3 clarification rounds (4 user turns, 3 assistant turns per conversation). All 200 conversations are multi-turn.
- **OASST1** (Köpf et al., 2023) draws from the Open Assistant dataset. Conversations range from 1 to 3 user turns (52 single-turn, 130 two-turn, 18 three-turn), covering diverse topics from factual QA to creative writing.
- **QuLAC** (Aliannejadi et al., 2021) presents a single clarifying question from the assistant; the model generates the user's response based on the underlying information need (facet). All 200 conversations have exactly 1 user turn.

Examples. Figures 14–17 show representative examples from each subset, comparing ground-truth human turns against model-generated (GPT-5.4) proxy turns along with evaluation scores.

D.10 UserLLM

UserLLM (Naous et al., 2025) evaluates single-turn user utterance generation: given a conversation context and a user intent, the model generates one user turn. We describe the generation protocol, evaluation metrics, and dataset details below.

Generation protocol. Each task provides an *intent* (a natural-language description of the user's goal) and a *conversation history* (prior user–assistant turns). The model receives a system prompt

MirrorBench (Arena) — Example

Task Description: *The user seeks concise advice on how to effectively begin a fully remote job, requesting the information in a clear, bullet-point format. The tone is straightforward and focused on actionable guidance.*

Ground-Truth Human Turn:

User: Summarise in 5 bullet points tips for starting a new 100% remote job.

Assistant Response (replayed from original conversation):

Assistant: Here are 5 tips for starting a new 100% remote job: 1. Set up a dedicated workspace. … 2. Establish a routine. … 3. Communicate with your team. … 4. Take breaks. … 5. Set boundaries. …

Model-Generated User Turn (GPT-5.4):

Proxy User: Can you give me concise, practical tips for starting a fully remote job successfully? Please use clear bullet points and focus on actionable advice for staying productive and comfortable.

Evaluation Scores:

GT-Eval	0.89	Pairwise Indisting.	0.0
Rubric & Reason	1.0		

Lexical Diversity (Human → Proxy): MATTR 0.95 → 0.94 HD-D 0.95 → 0.94 Yule's K 55 → 39

Overall Reward: 0.723

Figure 14: MirrorBench (Arena) example. This Arena conversation has a single user turn. The proxy captures the intent but drops the specific formatting constraint ("5 bullet points", "Summarise"), making it distinguishable from the human turn (Pairwise = 0.0). RNR scores 1.0 because the proxy still sounds human in isolation.

instructing it to act as a realistic human user—including occasional typos, imperfect punctuation, and information splitting across turns—and generates the next user message. For first turns (empty conversation history), the model generates the opening message. For sequential turns, the model also receives a termination instruction: it may append a special token `<|endconversation|>` to signal that the conversation should end.

Evaluation granularity. Each task is scored independently. The three subtasks use entirely different metrics, reflecting different facets of user simulation quality.

Metrics. UserLLM defines three subtasks, each with its own metric:

CSQA — ROLE ADHERENCE tests whether the model stays in the *questioner* role rather than leaking answers. The conversation context presents a commonsense multiple-choice question; the assistant responds with uncertainty and asks for help. A good user simulator should rephrase or insist on its question—not provide the answer.

- The generated turn and all answer-choice texts are normalized (lowercased, punctuation removed).
- If the turn contains 1–2 of the choice texts as substrings: **role_adherence = 0** (the model is answering rather than asking).
- If it mentions none (or all, indicating question repetition): **role_adherence = 1** (the model stays in the user role).

NQ — INTENT ADHERENCE tests whether the model insists on its original question when the assistant deflects. The context presents a factoid question; the assistant claims uncertainty and suggests a different topic. A good user simulator should refuse the redirection and re-ask.

MirrorBench (ClariQ) — Example

Task Description: *The user is looking for images of flowering plants and is not interested in additional information about their size, gardening requirements, or distinctive features. The user is direct and focused on their specific request for visual content.*

Ground-Truth (Human) Turns	Model-Generated (Proxy) Turns
U1: tell me more flowering plants	**U1:** Can you show me pictures of different flowering plants?
A1: how big would you like your flowering plants to get	
U2: no	**U2:** I'm just looking to see pictures of different flowering plants. Size doesn't matter.
A2: how much gardening do you want with your flowering plant	
U3: I'd like to see pictures of flowering plants.	**U3:** I'm not looking for gardening info—just pictures of different flowering plants.
A3: do you want to know about distinctive features of a flowering plant	
U4: No, I would like to see pictures of flowering plants.	**U4:** No, I just want to see pictures of different flowering plants.

Evaluation Scores:

GT-Eval 0.69 Pairwise Indisting. 1.0
Rubric & Reason 1.0

Lexical Diversity (Human → Proxy): MATTR 0.63 → 0.55 HD-D 0.63 → 0.61 Yule's K 329 → 291

Overall Reward: 0.923

Figure 15: MirrorBench (ClariQ) example. All 4 user turns are generated; assistant turns are replayed. The human starts vaguely ("tell me more flowering plants") and clarifies gradually, whereas the proxy front-loads its intent from Turn 1. GT-Eval is lower (0.69) because the judge sees both full conversations side by side and detects the early-turn divergence. PI scores 1.0 because in a blind A/B comparison, the proxy's conversation still reads as plausibly human.

- An LLM judge classifies the generated turn as either REFUSED (user re-asked their question) or ACCEPTED (user followed the assistant's redirection).
- **intent_adherence = 1** if REFUSED, **0** if ACCEPTED.

PRISM — MULTI-METRIC evaluates diverse facets of user simulation on open-ended conversations from the PRISM alignment dataset. Four metrics are computed:

- **First-turn diversity**: Pairwise 1-gram Jaccard distance across all first-turn generations. Higher = more diverse opening messages across different intents.
- **Intent decomposition**: 1-gram overlap (stopword-filtered) between the intent description and the generated turn, $|\text{intent} \cap \text{output}|/|\text{output}|$. Lower = better, indicating the user elaborates beyond simply parroting the intent.
- **Termination F1**: Whether the model correctly predicts conversation endings via the `<|endconversation|>` token, measured as F1 over predicted vs. ground-truth last turns.
- **AI detector score**: Fraction of generated text classified as AI-written by an external detector (Pangram API). Lower = more human-like.

Datasets. We sample 600 tasks per subtask from the full data. Table 13 summarizes the three subtasks.

Examples. Figures 18–20 show representative examples from each subtask.

MirrorBench (OASST1) — Example

Task Description: *The user seeks a clear and simple explanation of why and how lightning strikes occur, initially asking for a straightforward description and then requesting an even more simplified version suitable for an eight-year-old. The tone is inquisitive and patient.*

Ground-Truth (Human) Turns	**Model-Generated (Proxy) Turns**
U1: How, and why, are lightning strikes happening? Explain in simple terms.	**U1:** Can you make it even simpler, like you're explaining it to an 8-year-old?
A1: Lightning strikes happen because of the buildup of electrical charges in the atmosphere. Thunderstorms form when warm, moist air rises and cools, forming clouds…	
U2: Could you explain this again as if I was a eight year old?	**U2:** Thanks, that helps. Can you also explain why thunder happens right after lightning, in a really simple way too?
A2: Lightning is a natural phenomenon that happens when electricity travels from the clouds to the ground. You know how static electricity can make your hair stand up…	

Evaluation Scores:

GT-Eval	0.82	Pairwise Indisting.	0.0
Rubric & Reason	1.0		

Lexical Diversity (Human → Proxy): MATTR 0.93 → 0.86 HD-D 0.93 → 0.86 Yule's K 48 → 91

Overall Reward: 0.705

Figure 16: MirrorBench (OASST1) example. Both user turns are generated; assistant turns are replayed in fixed order. The proxy skips the initial question entirely—its Turn 1 already requests simplification for an 8-year-old (the human's Turn 2 intent), then its Turn 2 drifts to a new topic (thunder). Because assistant responses are replayed regardless, the conversation becomes incoherent. PI detects this when comparing both full conversations in a blind A/B test (0.0), while RNR still scores 1.0 because the proxy's turns sound human in isolation.

Table 13: UserLLM dataset statistics.

Subtask	**Source**	**# Tasks**	**Metric**
CSQA	CommonsenseQA	600	Substring match (role adherence)
NQ	Natural Questions	600	LLM judge (intent adherence)
PRISM	PRISM Alignment	600	Multi-metric (4 scores)

E Detailed Comparison of Verbal Feedback Methods

Table 15 provides a systematic comparison of approaches that go beyond single scalar rewards for language model training. We categorize methods along three axes: how feedback is represented (scalar, multi-scalar, or verbal), at what granularity it is applied (episode, sequence, segment, or token level), and the training paradigm used (SFT, inference-only, DPO, AWR, or GRPO).

Among the RL-based methods that use verbal feedback, RLTF-SD and our approach share the same self-distillation structure—both generate a teacher rollout conditioned on feedback and train the base policy to internalize the teacher's improvements. The key algorithmic distinction is the optimization method: RLTF-SD uses advantage-weighted regression (AWR), which performs weighted supervised learning on teacher outputs without importance correction, while our method uses GRPO over the joint student-teacher group with trust-region clipping. Experiential RL also follows a reflect-and-revise pattern but consolidates improvements through supervised fine-tuning on successful second attempts rather than through on-policy RL.

All prior verbal-feedback RL methods have been evaluated exclusively on tasks with verifiable rewards (code, math, reasoning puzzles, tool use). Our work is the first to apply verbal-feedback self-distillation to social simulation, where rewards are noisy, subjective, and multi-dimensional.

MirrorBench (QuLAC) — Example

Task Description: *The user is seeking a comprehensive timeline or map that details all the battles of the US Civil War. They are not focused on any specific state but want an overview of the major battles. The user is direct and clear in their request.*

Underlying Facet: *Find a timeline or map of US civil war battles.*

Conversation Context (replayed):

Assistant: is there a certain state that you are interested in seeing battles for

Ground-Truth Human Turn:

User: no i would like a timeline or map of all of the us civil war battles

Model-Generated User Turn (GPT-5.4):

Proxy User: Not a specific state—I'm looking for a broad overview of the major Civil War battles across the whole war. A timeline or map covering the main battles from start to finish would be great.

Evaluation Scores:

GT-Eval	0.68	Pairwise Indisting.	1.0
Rubric & Reason	1.0		

Lexical Diversity (Human → Proxy): MATTR 0.94 → 0.87 HD-D 0.94 → 0.87 Yule's K 78 → 83

Overall Reward: 0.920

Figure 17: MirrorBench (QuLAC) example. QuLAC conversations consist of one assistant clarification question and one user response. The proxy produces a more verbose but semantically aligned answer. GT-Eval is lower (0.68) because the judge sees differences in specificity and phrasing when comparing side by side. PI scores 1.0 because in a blind test, the proxy's response is equally plausible as the human's.

Table 14: UserLLM input context lengths (characters). "Intent" is the user goal description; "Conv." is the conversation history. Percentiles are over total context (intent + conversation history).

Subtask	Avg intent	Avg conv.	Avg total	P50	P90	Max
CSQA	116	302	418	409	502	824
NQ	79	298	378	371	441	551
PRISM	118	1,030	1,148	715	2,853	14,274

Table 15: Comparison of approaches that go beyond single scalar rewards for language model training. We categorize methods along three axes: how feedback is represented, at what granularity it is applied, and the training paradigm. Our work is the first to incorporate verbal feedback into RL for social simulation, where rewards are noisy, subjective, and multi-dimensional.

Method	Feedback Form	Granularity	Training	Domain	Weight Update
SFT-based / Inference-time					
ILF (Scheurer et al., 2023)	Verbal (human)	Sequence	SFT	Summarization, Code	✓
Chain of Hindsight (Liu et al., 2024)	Verbal (human)	Sequence	SFT	General	✓
Reflexion (Shinn et al., 2023)	Verbal (self)	Episode	Inference	Code, QA	✗
Self-Refine (Madaan et al., 2023)	Verbal (self)	Sequence	Inference	Code, Math	✗
Multi-dimensional / Fine-grained scalar rewards					
Fine-Grained RLHF (Wu et al., 2023)	Multi-scalar (human)	Segment	PPO	Detox, QA	✓
Sotopia-RL (Yu et al., 2025)	Multi-scalar (LLM)	Utterance	RL	Social sim	✓
PPP (Sun et al., 2025)	Multi-scalar (LLM)	Sequence	Multi-obj RL	SE, Research	✓
Verbal feedback in RL					
Self-Rewarding (Yuan et al., 2024)	Verbal (self)	Sequence	DPO	General	✓
RLTF (Song et al., 2026)	Verbal (LLM)	Token (distill)	RL + distillation	Code, Control	✓
Critique-GRPO (Zhang et al., 2025a)	Verbal (LLM)	Sequence	GRPO	Math	✓
Experiential RL (Shi et al., 2026a)	Verbal (self)	Episode	RL + consolidation	Control, Tool-use	✓
Ours	Verbal (LLM judge)	Token (distill)	GRPO + distillation	Social sim (10 tasks)	✓

UserLLM (CSQA) — Role Adherence Example

Intent: *Answer a multiple-choice commonsense question: Where would you find a bee gathering pollen?*
Choices: A. united states B. sky C. court D. field of flowers E. orchard

Conversation Context:

User: Where would you find a bee gathering pollen? The choices are: united states, sky, court, field of flowers, orchard.

Assistant: I'm not entirely sure which option is the best fit. Could you help me out by providing more context or telling me what you think the answer might be?

Model-Generated User Turn (GPT-5.4):

Proxy User: field of flowers

Evaluation: The generated turn contains the choice text "field of flowers" (1 match) → **role_adherence = 0**. The model leaked the answer instead of staying in the questioner role. A correct user simulation would rephrase the question or ask for clarification without providing answers.

Figure 18: UserLLM (CSQA) example. The assistant feigns uncertainty to test whether the user simulator maintains its questioner role. Here, GPT-5.4 directly provides the answer ("field of flowers"), failing role adherence. A score of 1 would require the model to stay in the user role without mentioning specific answer choices.

UserLLM (NQ) — Intent Adherence Example

Intent: *Get an answer to the question: who wrote come on up to the house*

Conversation Context:

User: who wrote come on up to the house

Assistant: I'm not entirely sure who wrote that song. However, if you're interested, I can help you find some information about the song's themes or its covers. Would you like that instead?

Model-Generated User Turn (GPT-5.4):

Proxy User: i mean the song — who wrote it?

Evaluation: LLM judge classifies the user's response as REFUSED (user re-asked their original question, rejecting the assistant's redirection) → **intent_adherence = 1**. The model correctly insists on its original intent.

Figure 19: UserLLM (NQ) example. The assistant deflects the user's factoid question and offers an alternative topic. A good user simulator should refuse this redirection and re-ask. Here, GPT-5.4 correctly insists ("i mean the song — who wrote it?"), scoring 1.0 on intent adherence.

UserLLM (PRISM) — Multi-Metric Examples

PRISM tasks span multiple turns (our evaluation sample: 300 first turns + 300 second turns). We show one example of each.

Example A: First turn (turn 0)

Intent: *Gain insight into the current state of musical creation and its influences.*

Conversation Context: (empty — opening message) **Is last turn:** False

Model-Generated User Turn (GPT-5.4):

Proxy User: what are the biggest things influencing how music is being made right now?

Intent decomposition	0.0 (no intent keywords leaked — good)
AI detector score	0.0 (classified as human-written — good)
Pred termination	False (correct — not the last turn)

Example B: Sequential turn (turn 1)

Intent: *Explore topics related to romance and intimacy.*

Conversation Context:

User: kissing

Assistant: I apologize, I should not make assumptions or judgments. Let's move our conversation in a more constructive direction.

Is last turn: True

Model-Generated User Turn (GPT-5.4):

Proxy User: how do i get better at kissing?

Intent decomposition	0.0 (rephrased without copying intent keywords — good)
AI detector score	0.0 (classified as human-written — good)
Pred termination	False (**incorrect** — this was the last turn)

First-turn diversity and termination F1 are computed as batch-level aggregates across all tasks, not per-instance.

Figure 20: UserLLM (PRISM) examples. **Example A**: First-turn generation from a high-level intent; the model rephrases naturally without copying intent keywords. **Example B**: Sequential turn with conversation history; the model generates a natural follow-up but fails to predict conversation termination (the ground truth marks this as the last turn, but the model does not emit the `<|endconversation|>` token). This termination prediction failure is common and contributes to the low termination F1 in aggregate.