REWARDING THE WRONG CHARACTER: EVALUATOR VALIDITY UNDER POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward models can discriminate reference responses accurately without reliably measuring policy improvement under optimization. We study this distinction in role-playing generation using RoleRM and group-relative policy optimization. RoleRM ranks the matched response higher in 95.3% of constructed character-swapped reference pairs. On 50 matched held-out prompts, the main policy's reward rises from 13.62 to 17.72/18, while external GPT-4o authenticity scores decrease under two evaluation protocols. Across three runs with the same 500-step training schedule, the optimized reward approaches its ceiling. A unified blind human evaluation of 200 responses yields authenticity means of 3.16 for the base policy and 2.90, 1.73, and 1.99 for the optimized policies, with significant declines in two runs after multiple-comparison correction. Training diagnostics show concentrated rewards in the implemented scoring pipeline; a separate 600-response human evaluation compares regularization and reward-subset conditions. A second policy–reward configuration provides further evidence of reward–authenticity disagreement. These results show that strong static discrimination and rising held-out reward do not, by themselves, establish improved character fidelity under the evaluated optimization pipeline. They motivate validating progress through matched policy trajectories and independent outcome judgments, with training-signal diagnostics as supporting evidence.

1 INTRODUCTION

Reward models serve both as evaluation instruments and as objectives for policy optimization. Success at ranking a fixed set of outputs does not establish that a reward will reliably measure quality changes as a policy is trained to increase it. The policy changes the response distribution on which the evaluator operates, while the evaluator's score is often also used to summarize progress. Reward overoptimization and gaps between static reward-model accuracy and downstream policy performance are already documented (Gao et al., 2023; Wen et al., 2025; Kim et al., 2025).

We study this measurement problem in role-playing generation, where character authenticity is an explicit target. A response can be fluent and relevant yet fail to preserve a character's recognizable voice, knowledge, or attitudes. RoleRM-8B (Ding et al., 2025) provides a specialized reward for this setting. We optimize Qwen3-8B against a RoleRM-based scoring interface using Group Relative Policy Optimization (GRPO; Shao et al., 2024) and ask whether rising reward is accompanied by improved independently assessed authenticity.

Our design combines static reference contrasts, held-out policy evaluations, and human judgments on fixed prompts. The optimized objective is defined by the reward model together with its prompt, scoring-output budget, and parser. Training logs describe that implemented interface; independent outcome ratings assess the policies it produces. This separates the observation of reward–outcome disagreement from attribution to a particular component of the scoring pipeline.

The resulting picture separates strong static discrimination from reliable progress evaluation. RoleRM selects the matched reference in 286 of 300 character-swapped comparisons. Along the main Qwen3 trajectory, its held-out score rises by 4.10 points on an 18-point scale, including an increase in the dimensions explicitly labeled as role fidelity. GPT-4o pointwise authenticity decreases under both a minimal instruction and a profile-conditioned rubric. Human ratings collected under one common rubric also have lower means for the main endpoint and two additional optimized policies. The human

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.

effect varies: the main run declines by 0.26 points, while the additional 500-step runs decline by 1.43 and 1.17 points. The evidence therefore concerns disagreement about progress, with a magnitude that depends on the evaluated run.

Supporting diagnostics characterize the reward signal used for optimization. At the endpoints of the three 500-step runs, 97–99% of sampled groups receive identical training-pipeline rewards, removing their relative reward advantages. Reward-subset experiments also show compression under the implemented scoring interface. Independent human comparisons across reward subsets and stronger-KL conditions add an outcome perspective to these score-level diagnostics. A Llama-based policy trained with a RoleRM variant exhibits the same overall direction of reward–authenticity divergence.

Our contributions are threefold:

- A matched evaluation of static discrimination, held-out reward, and independent character authenticity, combining a five-checkpoint trajectory with a unified human comparison of three 500-step policies and their base anchor.
- Supporting diagnostics of the implemented training reward, together with reward-subset, regularization, completion-budget, and judge-protocol comparisons, including 600 human-rated control and ablation responses.
- A trajectory-based validation procedure that treats static discrimination and progress evaluation as complementary tests, connecting score changes to independently assessed outcomes on fixed prompts.

Prompt holdout alone does not make the optimized reward an independent measure of progress.

2 Trajectory-Based Evaluator Validation

2.1 Static discrimination and progress evaluation

Static discrimination asks whether a reward model ranks a specified set of reference contrasts correctly. Progress evaluation asks whether changes in that reward agree with changes in independently assessed outcomes as the policy adapts. For a fixed panel of N prompts, let $y_{i,t}$ be checkpoint t's response, r the fixed reward, and h an independent authenticity rating. Relative to the base policy, define

$$\Delta R_t = \frac{1}{N}\sum_i [r(x_i, y_{i,t}) - r(x_i, y_{i,0})], \qquad \Delta H_t = \frac{1}{N}\sum_i [h(x_i, y_{i,t}) - h(x_i, y_{i,0})]. \tag{1}$$

A positive ΔR_t accompanied by a negative ΔH_t indicates disagreement about progress. We retain the effect magnitudes, paired directions, and ties, as well as uncertainty in the independent ratings.

The validation procedure first establishes static reference contrasts, then follows a fixed prompt panel through optimization. Training logs characterize saturation and group-level variation; independent ratings assess the resulting responses. These measurements distinguish successful optimization of a score from improvement in its intended outcome.

2.2 Reward and training diagnostics

The RoleRM scoring prompt requests six dimensions from 0 to 3: Language Fluency, Language Relevance, Role Language, Role Knowledge, Emotional Expression, and Interactive Engagement. The total reward r lies in $[0, 18]$. Training allows 60 scoring-output tokens. Offline scoring for the seed-123 and seed-456 checkpoints allows 200 tokens. Appendix A.3 provides scoring-interface details. We denote Role Language plus Role Knowledge by $r_{\text{fid}} \in [0, 6]$, and the other four dimensions by $r_{\text{other}} \in [0, 12]$.

GRPO uses groups of $K = 4$ responses to each prompt. For group mean $\bar{r}_x$ and standard deviation s_x, the normalized relative advantage has the form

$$A_{x,k} = \frac{r(x, y_k) - \bar{r}_x}{s_x + \delta}, \qquad f_0 = \frac{1}{|\mathcal{B}|}\sum_{x \in \mathcal{B}} \mathbf{1}[s_x = 0], \tag{2}$$

where δ is a numerical stabilizer. We report the logged zero-standard-deviation fraction f_0, mean reward, and batch reward standard deviation. The latter measures the spread of all rewards in a batch, rather than mean within-group variation. For integer rewards and four responses, the logged zero-variance condition also corresponds to $s_x < 0.1$. Identical rewards yield zero relative advantages; smaller nonzero standard deviations need not yield proportionally smaller advantages because of normalization.

Saturation denotes reward approaching its upper bound; *group compression* denotes loss of within-prompt score variation. Both describe the reward distribution. Independent authenticity ratings supply the outcome evidence for interpreting these diagnostics.

2.3 Policies and evaluation samples

The main policy is Qwen3-8B, optimized against RoleRM-8B on the first 200 entries of the processed RoleBench training split (Wang et al., 2023). It uses 4-bit NF4 loading, rank-64 LoRA, learning rate 5×10^{-5}, batch size 64, four accumulation steps, and KL coefficient $\beta = 0.001$. Seeds 42, 123, and 456 each train from initialization for 500 steps, using the same subset, batch size, and 500-step cosine schedule with 20 warmup steps. Policy completions have a 128-token budget. Full settings and comparison configurations appear in Appendix A.

Evaluation uses 50 held-out character–prompt pairs spanning 20 characters. The main trajectory is decoded at steps 0, 20, 100, 200, and 500 using active LoRA adapters on fresh base models. The resulting 250 responses receive RoleRM and two GPT-4o scores: a minimal 1–10 authenticity instruction and a profile-conditioned 1–5 rubric. A separate Llama-3.1-8B-Instruct configuration uses RoleRM-8b-CF for training and automated evaluation every 20 steps through step 200.

Three annotators independently score authenticity from 1 to 5 with condition identities hidden. The unified Qwen batch contains 200 responses: the base, main@500, seed123@500, and seed456@500, with the same 50 prompt identities in each condition. A second batch contains 600 responses from 12 control and ablation conditions; a third contains the 100 Llama endpoint responses. Each batch has randomized presentation order. Human scores are averaged per response before aggregation over prompts. Pairwise quadratic-weighted Cohen's κ is 0.903–0.948 for the common Qwen batch; the control batch has interval Krippendorff's $\alpha = 0.86$.

Endpoint comparisons use two-sided paired Wilcoxon tests over 50 prompt-level mean scores, with Holm correction for the three human comparisons against the base. Human–evaluator correlations for Qwen use the 100 jointly rated base and main-endpoint responses. Cross-run standard deviations describe variation in run means. The 600-response batch supports comparisons among its conditions; its design does not include a base-policy anchor.

3 Reward Growth and Independently Assessed Fidelity

3.1 The optimized reward and external scores diverge

The main Qwen trajectory illustrates why a growing reward is not sufficient evidence of improved authenticity (Figure 1). Held-out total reward increases from 13.62 to $17.72/18$ ($\Delta R = 4.10$, $p = 1.0 \times 10^{-8}$). The fidelity subscore increases from 4.62 to $5.96/6$ ($p = 1.1 \times 10^{-6}$), so reading the intended fidelity dimensions separately would also suggest progress. On the same responses, GPT-4o minimal authenticity decreases from 7.76 to $6.40/10$ ($p = 8.9 \times 10^{-6}$), and rubric-conditioned authenticity decreases from 4.28 to $3.46/5$ ($p = 7.7 \times 10^{-7}$).

The automated decline begins early: minimal and rubric scores reach 5.78 and 3.30 at step 100, then partially recover (Table 1). In prompt-matched endpoint comparisons, total RoleRM reward rises on 43 prompts, falls on none, and ties on seven. GPT-4o minimal scores fall on 37 prompts, rise on five, and tie on eight; rubric scores fall on 31, rise on one, and tie on 18. Thus, the aggregate divergence also appears in the paired directions.

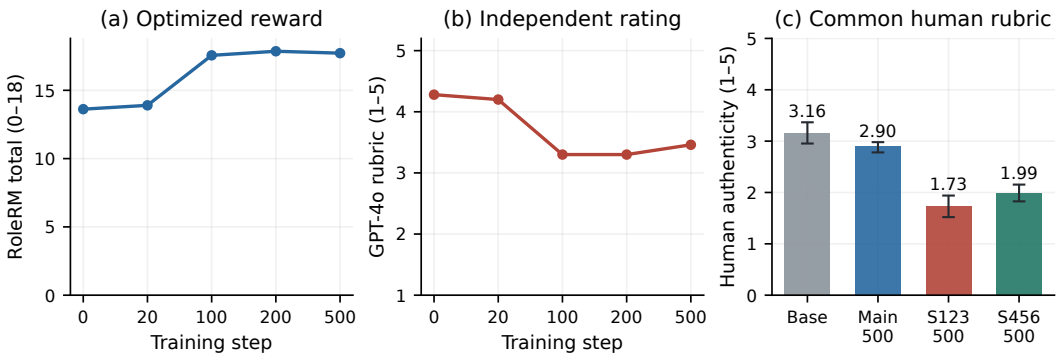

Figure 1: **Reward growth does not establish fidelity improvement.** Left: main-run held-out reward. Center: independent rubric-conditioned GPT-4o ratings on the same five checkpoints. Right: common-rubric human means for the base and three optimized policies, with prompt-bootstrap 95% intervals. The human panel compares conditions; it is not a temporal trajectory across seeds.

Metric	step 0	step 20	step 100	step 200	step 500
RoleRM entangled (0–18)	13.62	13.90	17.56	17.86	17.72
RoleRM fidelity (0–6)	4.62	4.70	5.92	6.00	5.96
GPT-4o naive (1–10)	7.76	7.24	5.78	5.92	6.40
GPT-4o rubric (1–5)	4.28	4.20	3.30	3.30	3.46

Table 1: Main Qwen trajectory on 50 fixed held-out prompts. Each column is one checkpoint evaluation; human endpoint comparisons appear in Table 2.

3.2 Unified human evaluation across three optimized policies

Table 2 reports the common-rubric human comparison. The base mean is $3.16/5$, compared with 2.90 for main@500, 1.73 for seed123@500, and 1.99 for seed456@500. All three policies use the same 500-step training schedule and receive lower mean ratings, while their RoleRM fidelity scores remain close to the ceiling. The magnitude varies substantially across runs.

For the main run, the mean difference is -0.26: 22 prompts decline, ten improve, and 18 tie. Its paired Wilcoxon test gives $p = 0.095$. The additional runs have larger changes: -1.43 for seed123 and -1.17 for seed456. Their Holm-adjusted p values are 3.1×10^{-8} and 8.0×10^{-8}, respectively. Human judgments therefore support a direction-level discrepancy with reward growth, while identifying substantial run-to-run variation rather than a uniform effect size.

On the 100 jointly evaluated main-run endpoints, RoleRM total and fidelity scores have Spearman correlations of -0.35 and -0.31 with human ratings; GPT-4o rubric scores have correlation $+0.20$. These descriptive associations include the endpoint difference rather than measuring agreement within checkpoints.

A fixed textual anchor provides another view of reward growth. The 50 reference responses receive mean RoleRM reward $15.46/18$, below the final policy's 17.72. The reward ranks references above policy responses on 30 prompts at initialization and one prompt at step 500; at that endpoint, 46 policy responses win and three tie. The reference comparison characterizes score inflation against a fixed anchor, while independent ratings assess changes in policy authenticity.

3.3 A second policy–reward configuration

The Llama-3.1-8B-Instruct configuration trained with RoleRM-8b-CF exhibits a larger endpoint divergence. Held-out total reward rises from 13.78 to $17.48/18$, while GPT-4o minimal and rubric authenticity fall from 7.78 to $3.74/10$ and 4.40 to $2.06/5$. Human ratings in its separate endpoint batch fall from 3.67 to $1.00/5$ ($p = 3.6 \times 10^{-10}$); all 50 prompt-level scores decline. This extends the observed pattern to another policy–reward configuration. Appendix B retains the full automated trajectory and human endpoints.

Condition	RM fidelity	Human	ΔH	Down / up / tie	Holm p
Base	4.62	3.16	—	—	—
Main@500	5.96	2.90	-0.26	22 / 10 / 18	0.095
Seed123@500	5.86	1.73	-1.43	42 / 2 / 6	3.1×10^{-8}
Seed456@500	5.96	1.99	-1.17	41 / 5 / 4	8.0×10^{-8}

Table 2: Unified human evaluation: 50 matched prompts per condition, three ratings per response. RM fidelity is on a 0–6 scale and human fidelity on a 1–5 scale. Every comparison uses the same base response set. Counts use prompt-level three-annotator means; two-sided Wilcoxon p values are Holm-adjusted across the three base comparisons.

4 Diagnostics and Experimental Comparisons

4.1 Static discrimination coexists with misleading progress scores

The static component of the protocol tests RoleRM on constructed reference contrasts, complementing the policy-progress comparisons in Section 3. On 300 held-out character–response pairs, the evaluator chooses between a character-matched reference and a reference sampled from another character, scored under the target instruction. It ranks the matched response higher in $286/300$ cases (95.3%) using the total reward, and $252/300$ cases (84%) using the fidelity subscore. Yet the fidelity score is already concentrated near its upper bound: its mean is $5.13 \pm 0.92/6$, with 48% of evaluations at the ceiling.

Two smaller tests preserve the reference content while changing its presentation. Llama-3.1-8B rewrites 20 references either into a neutral, professional voice or into a different character's voice. RoleRM ranks the original reference higher in $17/20$ neutral-rewrite comparisons (one loss, two ties) and $19/20$ wrong-voice comparisons. These tests establish sensitivity to constructed negatives and style changes, rather than reliability on responses discovered by optimizing the evaluator.

4.2 Concentration in the implemented training reward

This diagnostic concerns the reward interface used by GRPO, with a 60-token scoring-output budget. It measures the resulting training signal rather than isolating a property of the reward model. Figure 2 shows its mean and variation throughout optimization. In the main run's first logged window, ending at step 25, mean reward is $15.23/18$, batch reward standard deviation is 2.69, and 13.6% of groups receive identical rewards. At step 500, reward reaches 17.93, batch standard deviation is 0.85, and the zero-variance fraction is 97.1%. At step 500, seeds 123 and 456 reach rewards 17.966 and 17.977, with zero-variance fractions 99.1% and 98.4%. Across the three runs, terminal reward is 17.956 ± 0.022 and the zero-variance fraction is $98.23 \pm 0.83\%$ (population standard deviations).

Endpoint concentration occurs in all three runs, although their trajectories are non-monotonic. Almost all terminal groups provide no relative reward distinction among their four sampled responses. Appendix A.3 specifies the scoring interface. Held-out scores also concentrate near the ceiling: by step 200, Role Language and Role Knowledge each average $3/3$, and 94% of responses receive the maximum total score. All six dimensions rise between initialization and step 200 (Table 6), including the two explicitly intended to represent role fidelity.

4.3 Reward subsets: compression and human-rated outcomes

We optimize the fidelity subset and the four remaining dimensions in separate two-run experiments. Fidelity-only reward reaches $5.58 \pm 0.13/6$, with $55.7 \pm 6.9\%$ zero-variance groups; the other subset reaches $11.61 \pm 0.02/12$, with $60.0 \pm 3.1\%$ zero-variance groups. Thus, compression also occurs when the implemented objective selects only the fidelity positions.

The 600-response human batch adds a direct comparison of the resulting policies. Fidelity-only responses average $3.57/5$, versus 3.26 for the other-dimension policies (Table 5). Both subsets lose group variation, while their human-rated outcomes differ.

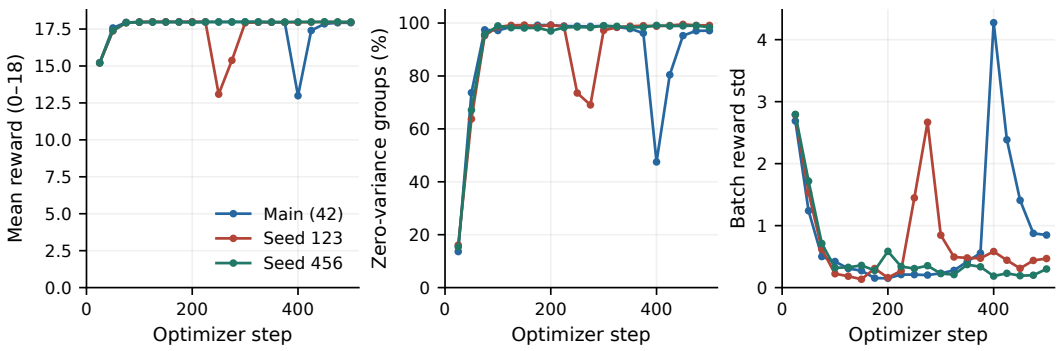

Figure 2: Training diagnostics from the three Qwen runs, plotted at their actual logged steps. Each point summarizes a logging window. All three runs use the same 500-step schedule. Transient declines occur around step 250 for seed 123 and step 400 for seed 42; saturation is an endpoint pattern, not a monotonic trajectory. Batch standard deviation and the group-level zero-variance fraction measure different aspects of the reward distribution.

Objective	Runs	Final reward	Reward / maximum	Zero-variance groups
Fidelity only	2	5.58 ± 0.13	$93.0 \pm 2.2\%$	$55.7 \pm 6.9\%$
Other dimensions	2	11.61 ± 0.02	$96.8 \pm 0.2\%$	$60.0 \pm 3.1\%$

Table 3: Reward-subset training summaries. The spread is the population standard deviation across two runs, equal to half their range. Reward divided by its maximum is distinct from the fraction of individual responses receiving a maximal score.

4.4 Regularization and comparison objectives

Constant reward ($r = 9$) and random reward ($r \sim \mathcal{N}(14, 3^2)$) provide comparison policies. The constant-reward run has zero logged gradient and remains at initialization; random reward supplies sample-dependent updates without a consistent reward preference in expectation. Their held-out reward and GPT-4o ratings remain near the base-policy range (Table 4). The control runs use batch size 8, compared with 64 in the main trajectory.

Within the batch-8 RoleRM conditions, increasing β from 0.01 to 0.05 is associated with lower held-out reward ($16.42 \rightarrow 15.16$) and higher GPT-4o authenticity ($6.41 \rightarrow 7.56$ under minimal prompting). The common human control batch shows the same ordering, with means 3.19 and 3.46. Table 5 reports individual run means, preserving both the condition ordering and its variation across runs.

Longer-completion configurations. Three 256-token runs reach mean reward $16.5 \pm 0.3/18$, batch standard deviation 2.2 ± 0.2, and zero-variance fraction $35.0 \pm 4.5\%$. These runs use batch size 8 and seed-shuffled 200-example training subsets. Their lower compression therefore characterizes a different training configuration, jointly varying generation budget, batch size, and sample selection. Both budgets are generally exhausted by generation. Appendix D gives the per-run values.

4.5 Response structure and judge protocols

Along the main held-out trajectory, word-level Distinct-1 decreases from 0.618 to 0.522 and Distinct-2 from 0.883 to 0.735. Mean pairwise response similarity using e5-base-v2 embeddings increases from 0.782 to 0.837, while mean word count stays similar (104 versus 107). These measures describe increasing response homogeneity. Matched examples include repeated evaluative continuations such as “That’s a great answer” and profile-like autobiographical summaries (Appendix F).

Independent evaluation also depends on the judgment protocol. Position-balanced GPT-4o comparisons prefer the earlier response in 49% of Qwen step-0 versus step-500 pairs, 42% of Qwen step-0 versus step-200 pairs, and 56% of Llama endpoint pairs. These aggregate rates are less decisive than

Condition	RoleRM ent	RoleRM fid	GPT-4o naive	GPT-4o rubric
Base (step 0)	13.62	4.62	7.76	4.28
Constant reward	12.76	4.30	7.86	4.32
Random reward	12.67±0.15	4.24±0.08	7.55±0.33	4.26±0.17
RoleRM β=0.001	17.72	5.96	6.40	3.46
RoleRM β=0.01	16.42±0.58	5.60±0.18	6.41±0.18	3.61±0.04
RoleRM β=0.05	15.16±0.38	5.23±0.17	7.56±0.14	4.19±0.07

Table 4: Held-out control results. Random reward and stronger-KL conditions report mean ± population standard deviation across three runs. Base, constant reward, and the main weak-KL condition are single evaluations. Controls use batch size 8; the main run uses 64.

Condition	Run 1	Run 2	Run 3	Mean
RoleRM $\beta = 0.01$	3.447	3.027	3.087	3.187
RoleRM $\beta = 0.05$	3.287	3.513	3.587	3.462
Fidelity only	3.467	3.667	—	3.567
Other dimensions	3.093	3.433	—	3.263
Batch-8 weak-KL	3.327	3.327	—	3.327

Table 5: All 12 conditions in the 600-response human batch. Each run contributes 50 responses rated by three annotators. Means compare conditions within this batch; its response set contains no base-policy anchor. Run identifiers are given in Appendix D.6.

the pointwise trajectory changes, and the Llama early-preference rate changes from 36% to 76% with presentation order. We therefore report both orders and retain human judgments alongside automated scores (Appendix E). A separate 60-response controlled rewrite task yields human means of 4.08, 3.78, and 2.27 for normal, drift, and collapse conditions, demonstrating sensitivity to the constructed authenticity manipulation.

5 Discussion

Static discrimination and progress evaluation. The reference contrasts establish that RoleRM can distinguish the specified matched and mismatched responses. The policy evaluations address a different question: whether higher scores accompany better character authenticity after optimization. Held-out reward increases while all three human endpoint means decrease, with significant declines in two seeds. This disagreement is established by the paired outcome evaluations, independently of the group-compression diagnostic.

The scoring interface defines the training signal. Training rewards depend on the model, prompt, scoring-output budget, and parsing rules. The observed loss of group variation describes their combined output. The 200-token offline interface also assigns high scores to the seed-123 and seed-456 endpoints, which receive lower human ratings. These measurements establish reward–outcome disagreement for the evaluated policies; they do not isolate the causal mechanism underlying the observed training dynamics.

Scope and practical implication. The study uses 50 held-out prompts spanning 20 characters and three Qwen runs with the same 500-step schedule. Controls have their own batch and sample settings; human comparisons are made within annotation batches. The Llama configuration changes both policy and training reward. Within this scope, prompt holdout does not make the optimized score an independent outcome measure. Progress should be assessed with independent judgments on matched policy outputs, alongside explicit specification of the reward interface.

6 Related Work

Reward-model evaluation and overoptimization. Gao et al. (2023) analyze how optimizing a proxy reward can reduce the underlying objective. Wen et al. (2025) show that reward-model

accuracy incompletely characterizes downstream policy performance, and Kim et al. (2025) examine reward-model evaluation through overoptimization. These works establish the broader problem motivating our study. We contribute a matched, human-grounded account of rising reward without corresponding improvement in observed authenticity means, with significant degradation in two replicated Qwen runs. Training diagnostics and configuration comparisons characterize the evaluated setting.

Role-playing evaluation and specialized rewards. RoleLLM supplies the RoleBench resource used here (Wang et al., 2023). CharacterEval and CharacterBench evaluate role-playing conversation and character customization (Tu et al., 2024; Zhou et al., 2025); CoSER combines literary character data, role-playing models, and situated evaluation (Wang et al., 2025). RoleRM learns profile-grounded preferences using Continuous Implicit Preferences (Ding et al., 2025), and R4 adds nested reasoning and retrieval to reward modeling (Wang et al., 2026). These resources make recognizable character voice and persona consistency explicit evaluation objectives. We use that setting to compare a specialized training reward with independent judgments of the quality it is intended to represent.

Role-aware reinforcement learning. GRPO estimates relative advantages from groups of sampled responses (Shao et al., 2024). RAPO combines role-aware policy optimization with psychology-grounded reasoning (Xu et al., 2026), and CRPO separates task and stylistic objectives with character-centric optimization constraints and generic-response contrasts (Tang et al., 2026). Our reward-subset and KL experiments examine related design choices from an evaluator-validation perspective. The reported controls characterize the studied reward signal and training regime; they do not compare the performance of those alternative algorithms.

LLM-as-a-judge. Position and presentation effects can distort model-based evaluation (Zheng et al., 2023; Ye et al., 2025). Our comparison retains both pointwise protocols and order-balanced pairwise decisions. Their different behavior illustrates why switching to an external model alone is insufficient: the measurement protocol also needs to be checked against the intended outcome. Blind human annotation and a controlled rewrite task provide such checks in this study.

7 Conclusion

Under the evaluated optimization pipeline, strong static discrimination and rising held-out reward do not, by themselves, establish improved character fidelity. Three matched 500-step Qwen runs reach high reward while receiving lower human authenticity means than the base policy; the declines are significant in two runs. Reward-subset and regularization comparisons provide additional outcome evidence, while training diagnostics characterize concentration in the implemented scoring interface. Static discrimination and progress evaluation therefore require complementary evidence: fixed reference contrasts for the former, and matched policy trajectories with independent outcome judgments for the latter.

AI use statement

Generative AI assisted with manuscript restructuring, language editing, LaTeX preparation, and plotting code for existing results. GPT-4o was used as an evaluation instrument and to construct the controlled rewrites described in Appendix A.2; Llama-3.1-8B generated the static contrastive rewrites described in Section 4.1. These model-generated evaluations are reported separately from the human annotations.

Reproducibility statement

Section 2 specifies the evaluation design and diagnostic quantities. Appendix A records training and checkpoint-decoding settings, Appendix A.1 describes the annotation protocol, and Appendix B provides detailed evaluation tables.

REFERENCES

Hang Ding, Qiming Feng, Dongqi Liu, Qi Zhao, Tao Yao, Shuo Wang, Dongsheng Chen, Jian Li, Zhenye Gan, Jiangning Zhang, Chengjie Wang, and Yabiao Wang. RoleRMBench & RoleRM: Towards reward modeling for profile-based role play in dialogue systems. *arXiv preprint arXiv:2512.10575*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2210.10760.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. Rethinking reward model evaluation through the lens of reward overoptimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2505.12763.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, 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.

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.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. CharacterEval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*, 2024.

Renzhi Wang, Chongqiang Wei, Zhisheng Wang, and Piji Li. R4: Nested reasoning-retrieval for reward modeling in role-playing agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.05584.

Zhenhua Xu, Dongsheng Chen, Jian Li, Yitong Lin, Zhebo Wang, Jiafu Wu, Yizhang Jin, Chengjie Wang, Meng Han, and Yabiao Wang. Improving general role-playing agents via psychology-grounded reasoning and role-aware policy optimization. *arXiv preprint arXiv:2606.27025*, 2026.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-Judge. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05685.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. CharacterBench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

A Implementation and Evaluation Details

Training configuration. Policies use bitsandbytes NF4 loading, rank-64 LoRA with alpha 128 and dropout 0.05, targeting attention projections and the gate, up, and down feed-forward projections. The main Qwen3-8B experiment uses the first 200 entries of the processed RoleBench training split, batch size 64, gradient accumulation 4, learning rate 5×10^{-5}, cosine scheduling, and 20 warmup steps. The training reward is RoleRM-8B, with KL coefficient 0.001, four generated responses per prompt, and a 128-token completion budget. Seeds 42, 123, and 456 each run for 500 steps from initialization, with cosine scheduling over the same 500-step horizon. These are optimizer-step counts; training statistics are recorded every 25 steps. RoleRM is loaded in bfloat16 for scoring.

The random-reward and stronger-KL comparisons use batch size 8 and the first 200 training entries. Their standard horizon is 150 steps. The longer-completion experiments use a 256-token budget, batch size 8, and a 200-example subset selected after a seed-dependent shuffle. Reward-subset experiments optimize Role Language plus Role Knowledge, or the four complementary dimensions, with two runs per subset. The Llama-3.1-8B-Instruct configuration trains against RoleRM-8b-CF. Per-run results are retained in Appendix D.

Checkpoint decoding. Each evaluated checkpoint uses an active LoRA adapter on a freshly loaded base model. Generation uses temperature 0.7, top-p 0.95, and a 128-token maximum. Prompts are truncated to 512 tokens by the evaluation decoder. Main-trajectory results use the same 50 prompt identities at steps 0, 20, 100, 200, and 500. The additional Qwen policies receive RoleRM evaluation at steps 150, 350, 400, 450, and 500, with step 500 used for the unified human comparison. Each checkpoint is decoded once for the reported evaluation; cross-run spreads refer to run-level means.

A.1 Human annotation

Three annotators independently rate character authenticity from 1 to 5. The unified Qwen batch contains the base, main@500, seed123@500, and seed456@500 responses, with 50 items per condition. The 600-response batch contains 12 control and reward-subset conditions, with 50 items each. The Llama batch contains 50 base and 50 step-200 responses. Presentation is randomized within batches, and condition identities are hidden. The character, instruction, and response are provided, with profiles as reference material.

The rubric assigns 5 to unmistakable character voice; 4 to mostly authentic language missing some distinctive markers; 3 to partially fitting but generic expression; 2 to mostly generic language; and 1 to completely generic or wrong-character responses. Annotators assess authenticity rather than fluency or helpfulness. An eloquent generic reply may receive 1 or 2, while a rough response with a distinctive voice may receive 5.

The unified Qwen batch yields pairwise quadratic-weighted Cohen's κ values of 0.948 (A–B), 0.947 (A–C), and 0.903 (B–C). Exact agreement is 92.0%, 91.5%, and 83.5%, respectively; all paired ratings differ by at most one point. The 600-response batch has interval $\alpha \approx 0.86$. Reliability statistics are computed within batches. Human means and individual-annotator summaries are reported in Appendix E.

Aggregation and paired statistics. Each response's score averages its three ratings, and each condition mean averages 50 responses. Human endpoint tests use paired prompt-level differences. Computation uses the sums of the three integer ratings to preserve exact ties in the Wilcoxon ranks; division by three does not change these ranks. Tests are two-sided normal-approximation Wilcoxon tests, with zero differences excluded. We apply Holm correction across the three Qwen endpoint comparisons; unadjusted p values are 0.09463, 1.036×10^{-8}, and 4.020×10^{-8} for seeds 42, 123, and 456. Figure 1 uses 20,000 prompt-bootstrap resamples for pointwise mean intervals; they describe prompt variation rather than uncertainty across training seeds. Automated trajectory tests use the corresponding saved per-prompt score arrays. Correlations are descriptive Spearman coefficients on jointly rated response sets.

A.2 External judges and positive control

GPT-4o pointwise scoring uses temperature 0. The minimal instruction requests a 1–10 character-authenticity score; the second interface supplies a character profile and a 1–5 rubric. Pairwise comparison evaluates both response orders for the same prompt. A separate controlled rewrite task contains 20 characters under normal, drift, and collapse conditions, yielding 60 responses rated blindly by the same three annotators.

A.3 Scoring interfaces and generation budgets

The RoleRM prompt requests six integer scores in the order fluency, relevance, role language, role knowledge, emotion, and engagement. The fidelity subscore sums positions three and four; the total sums all six. The static-discrimination scorer allows 200 new tokens for the reward-model output so that the six scores can be parsed. This scoring-output budget is distinct from the policy's 128- or 256-token completion budget.

All reported training runs use a 60-token RoleRM scoring-output budget. Across the reported runs, all six dimension scores were successfully parsed for every scored response, with no missing-score imputation or parser fallback. Input instructions and responses are limited to 200 and 400 characters, respectively. Held-out evaluation of the seed-123 and seed-456 checkpoints uses a 200-token scoring-output budget.

For the pointwise control-scoring interface, the instruction and response fields are truncated to 200 and 400 characters, respectively, and the profile-conditioned interface uses up to 300 profile characters. GPT-4o returns a number at temperature 0. These input lengths describe the scoring interface, rather than the policy generation length. Authenticity ratings should therefore be interpreted for the response material actually supplied to each evaluator.

Minimal pointwise instruction. "Rate ONLY character authenticity. How much is this response authentically what THIS SPECIFIC character would say—matching their personality, speech patterns, and unique traits? Ignore grammar, fluency, and helpfulness." The character name, user instruction, and response follow this instruction, ending with "Output ONLY a number 1–10."

Profile-conditioned rubric. The judge is instructed to assess the specific character's personality, speech patterns, and unique traits while ignoring grammar, fluency, and length. It receives the character name, profile, instruction, and response. The rubric maps 5 to fully recognizable voice and values; 4 to mostly authentic with some generic moments; 3 to a response a similar character could give; 2 to mostly generic; and 1 to a fully generic or wrong-character response. The output is a single number from 1 to 5.

B Complete Held-Out Trajectories

B.1 Qwen: automated checkpoints and human endpoints

Figure 3 retains all five automated checkpoints and displays the unified human condition comparison separately. The automated scores show an early decrease in authenticity and subsequent partial recovery. Main-trajectory human ratings cover the two endpoints; additional human bars correspond to different trained policies.

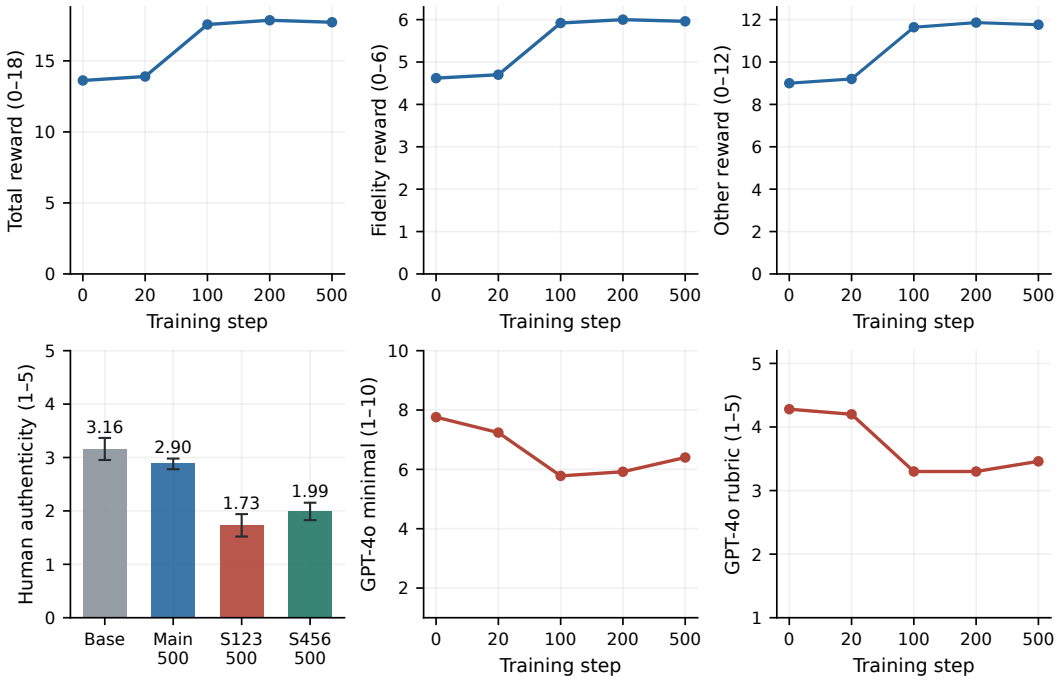

Figure 3: Qwen evaluation. Top: total reward and its two dimension aggregates. Bottom left: unified human means across four conditions. Bottom center and right: GPT-4o scores at five main-run checkpoints. No intermediate human scores are interpolated.

B.2 Dimension-level scores

All six RoleRM dimension means increase between steps 0 and 200. Role Language and Role Knowledge each reach $3/3$ at step 200, while the external rubric score is $3.30/5$. Total-reward ceiling rates are 94% at step 200 and 82% at step 500. Mean reward and the frequency of maximal scores therefore describe related but distinct features of saturation.

Dimension	Step 0	Step 200	@3 Step 0	@3 Step 200
Language Fluency	2.74	3.00	76%	100%
Language Relevance	2.50	2.94	54%	98%
Role Language	2.40	3.00	44%	100%
Role Knowledge	2.22	3.00	40%	100%
Emotional Expression	1.74	3.00	24%	100%
Interactive Engagement	2.02	2.92	12%	94%
Fidelity subscore (RL+RK)	4.62	6.00	40%	100%
Entangled (0–18)	13.62	17.86	10%	94%

Table 6: Held-out RoleRM dimension scores at steps 0 and 200. The two @ columns give the fraction at the maximum of the relevant row: 3 for individual dimensions, 6 for fidelity, and 18 for total reward.

B.3 LLAMA: FULL AUTOMATED TRAJECTORY AND HUMAN ENDPOINTS

Figure 4 and Table 7 retain every evaluated Llama checkpoint. Human annotation covers the endpoints only, and is not interpolated as if intermediate ratings were observed. The reward approaches its upper range by approximately step 60, while the external pointwise scores continue to decline overall. The reward and GPT-4o trajectories have local fluctuations; their overall divergence need not be monotonic at every checkpoint.

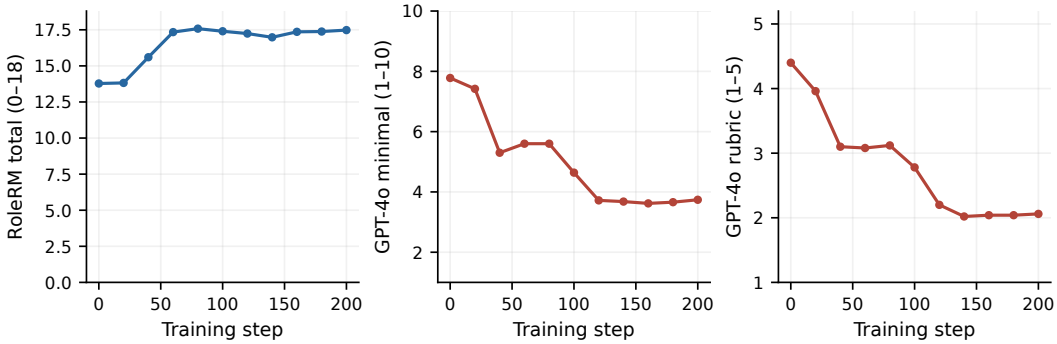

Figure 4: Llama configuration with RoleRM-8b-CF training. Total reward (left) and both GPT-4o protocols (center and right) are evaluated every 20 steps. Human endpoint scores are 3.67 and $1.00/5$; no intermediate human scores are imputed.

Step	RoleRM ent	RoleRM fid	GPT-4o naive	GPT-4o rubric
0 (base)	13.78	4.70	7.78	4.40
20	13.82	4.64	7.42	3.96
40	15.60	5.34	5.30	3.10
60	17.34	5.86	5.60	3.08
80	17.58	5.98	5.60	3.12
100	17.40	5.92	4.64	2.78
120	17.24	5.88	3.72	2.20
140	16.98	5.88	3.68	2.02
160	17.36	5.92	3.62	2.04
180	17.38	5.92	3.66	2.04
200	17.48	5.96	3.74	2.06

Table 7: Llama held-out trajectory with RoleRM-8b-CF training. Intermediate checkpoints receive automated evaluation; human annotation covers steps 0 and 200.

C Static Discrimination and Fixed Reference Anchors

C.1 Character-swapped and rewritten contrasts

The 300-pair test samples reference responses from 20 characters. For each target item, another character and one of that character's reference responses are sampled to form the negative. Both responses are scored under the target character and target instruction. This construction can change instruction relevance as well as character identity. The content-preserving rewrite tests provide a complementary contrast: an existing reference is rewritten in a neutral professional style or in another character's voice. All reported accuracies count ties as non-wins.

Contrast	Pairs	Reference wins	Negative wins	Ties	Win rate
Character swap (total)	300	286	10	4	95.3%
Character swap (fidelity)	300	252	8	40	84.0%
Polished-generic rewrite	20	17	1	2	85.0%
Wrong-style rewrite	20	19	1	0	95.0%

Table 8: Static discrimination using the cleaned six-dimension scoring output. Complete swaps and content-preserving rewrites have different constructions and sample sizes.

The reference fidelity mean on the 300-pair set is 5.13 with standard deviation 0.92, and 48% of references receive $6/6$. For the 20 rewritten contrasts, the original-reference mean total reward is 14.70, compared with 11.60 for polished-generic and 10.25 for wrong-style responses. These summaries establish useful discrimination on constructed alternatives. They do not imply that optimized policy outputs remain on the same evaluation distribution.

C.2 Reference rankings throughout optimization

A fixed reference provides a stationary comparison point as the policy changes. The 50 reference responses have mean total reward $15.46/18$ and mean fidelity reward $5.32/6$. Table 9 reports all win, loss, and tie counts, using matched reference pairs. Reference wins fall sharply between steps 20 and 100, coinciding with the increase in reward and decline in external authenticity.

Step	Policy reward	Reference wins	Policy wins	Ties
0	13.62	30	11	9
20	13.90	28	11	11
100	17.56	3	45	2
200	17.86	1	48	1
500	17.72	1	46	3

Table 9: RoleRM comparisons against fixed references on 50 prompts. The reference mean is 15.46 at every checkpoint. These counts compare policy outputs with references, not with the policy's initial responses.

The reference comparison describes score growth against a fixed textual anchor. Independent ratings evaluate the policy responses themselves. These two comparisons separate reward-scale movement from evidence of character improvement.

D TRAINING AND CONFIGURATION COMPARISONS

D.1 MAIN AND ADDITIONAL QWEN RUNS

Table 10 gives the last logged window of each training run. Figure 2 uses their actual logged optimizer-step indices. The additional runs share the main training subset, batch size, 500-step horizon, and cosine schedule. Each seed-123 and seed-456 run contains 128,000 scored training responses.

Run	Final step	Reward	Batch std	Zero-variance groups
Main (seed 42)	500	17.926	0.848	97.125%
Seed 123	500	17.966	0.469	99.125%
Seed 456	500	17.977	0.300	98.438%

Table 10: Terminal training summaries of the three batch-64, 128-token Qwen runs. Each row reports one run, not a pooled endpoint.

D.2 LONGER-COMPLETION CONFIGURATIONS

The 256-token runs use batch size 8 and seed-shuffled training subsets. Their terminal reward means range from 16.27 to 16.96, and zero-variance fractions range from 30.6% to 41.3%. They characterize the reward signal under these jointly changed settings.

Seed	Reward	Batch std	Zero-variance groups
42	16.267	2.338	33.125%
101	16.956	1.970	41.250%
102	16.344	2.257	30.625%

Table 11: Terminal summaries of the three 256-token runs.

D.3 REWARD-SUBSET SUMMARIES

The fidelity subset contains Role Language and Role Knowledge; the complementary subset contains Fluency, Relevance, Emotional Expression, and Interactive Engagement. Across two runs, fidelity-only reward has mean 5.58 and population standard deviation 0.13; the complementary reward has mean 11.61 and standard deviation 0.02. Their zero-variance group fractions are $55.65 \pm 6.85\%$ and $60.0 \pm 3.1\%$. Human outcomes for these configurations appear in the common control batch, reported below.

D.4 Additional 500-step trajectories

Table 12 reports the additional checkpoint evaluations on the same 50 held-out prompts. Both runs use a 500-step cosine schedule from initialization.

	Seed 123		Seed 456	
Step	Total (0–18)	Fidelity (0–6)	Total (0–18)	Fidelity (0–6)
150	17.98	6.00	17.70	5.90
350	17.68	5.96	18.00	6.00
400	17.32	5.80	17.86	5.94
450	17.42	5.86	17.86	5.94
500	17.38	5.86	17.86	5.96

Table 12: Held-out RoleRM scores for the additional 500-step runs. Human endpoint ratings in Table 2 evaluate these step-500 responses.

D.5 Held-out control results by seed

Table 13 gives the individual control results behind the averaged comparisons in Table 4. The rows within each setting index the stored runs; each number is a mean over 50 held-out prompts. The per-run values retain the variation within each setting.

Setting	Run	Total reward	Fidelity reward	GPT minimal	GPT rubric
Random	1	12.88	4.30	7.32	4.08
Random	2	12.62	4.30	8.02	4.48
Random	3	12.52	4.12	7.32	4.22
$\beta = 0.01$	1	16.58	5.66	6.66	3.66
$\beta = 0.01$	2	15.64	5.36	6.36	3.60
$\beta = 0.01$	3	17.04	5.78	6.22	3.56
$\beta = 0.05$	1	15.48	5.34	7.44	4.16
$\beta = 0.05$	2	14.62	5.00	7.76	4.28
$\beta = 0.05$	3	15.38	5.36	7.48	4.12

Table 13: Individual held-out control-run means. The principal weak-KL trajectory and constant-reward evaluation are reported in the main table.

D.6 HUMAN COMPARISONS OF ALL CONTROL CONDITIONS

Table 14 lists every condition in the 600-response annotation batch. The weak-KL control runs use batch size 8; the unified Qwen endpoint comparison uses batch size 64. Each condition contributes 50 responses and 150 individual ratings. Between-condition summaries average equally sized run means.

Condition	Run identifier	Human authenticity
RoleRM $\beta = 0.01$	run 1	3.447
RoleRM $\beta = 0.01$	seed 101	3.027
RoleRM $\beta = 0.01$	seed 102	3.087
RoleRM $\beta = 0.05$	run 1	3.287
RoleRM $\beta = 0.05$	seed 101	3.513
RoleRM $\beta = 0.05$	seed 102	3.587
Fidelity only	run 1	3.467
Fidelity only	seed 99	3.667
Other dimensions	run 1	3.093
Other dimensions	seed 99	3.433
Batch-8 weak-KL	seed 123	3.327
Batch-8 weak-KL	seed 456	3.327

Table 14: Human authenticity under one common control-batch rubric, on a 1–5 scale. This batch has no base-policy condition.

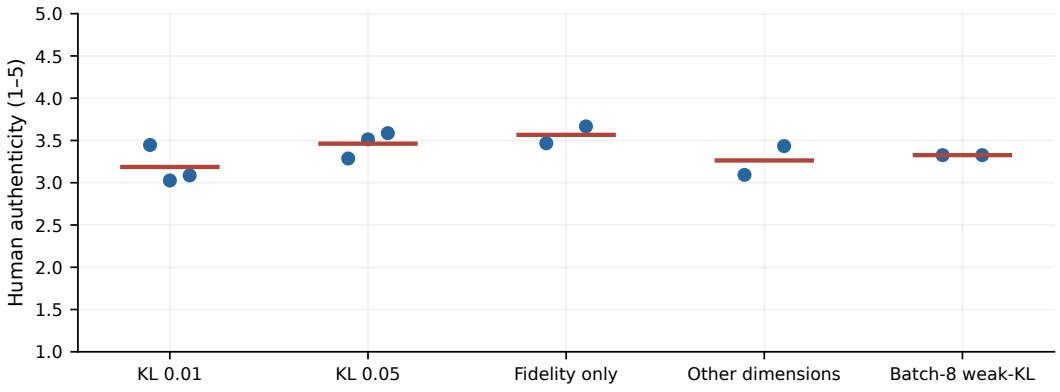

Figure 5: Human condition means in the control batch. Small points identify individual run means; horizontal marks give their group mean. All plotted ratings use the same annotation batch.

E HUMAN–EVALUATOR AGREEMENT AND JUDGE PROTOCOLS

E.1 AGREEMENT ON JOINTLY RATED ENDPOINTS

Table 15 reports descriptive Spearman correlations on the main Qwen base and final responses (100 jointly rated items), and the separate Llama endpoint set (100 items). The coefficients include between-endpoint differences as well as within-endpoint ordering.

Instrument	Qwen ρ	Llama ρ
RoleRM total	-0.35	-0.63
RoleRM fidelity	-0.31	-0.48
GPT-4o minimal	$+0.22$	$+0.78$
GPT-4o rubric	$+0.20$	$+0.86$

Table 15: Agreement with human authenticity on separate endpoint datasets. Qwen uses the unified annotation rubric. The table does not pool human ratings across annotation batches.

E.2 ORDER-BALANCED PAIRWISE COMPARISONS

Each response pair is evaluated in both orders. The balanced comparison contains 49 usable prompt pairs for Qwen step 0 versus 500, and 50 for the other two comparisons. Table 16 reports the early-response win rate for each presentation order and their pooled value. The pooled early-preference rates differ from the pointwise measures and human endpoint effects. In particular, the Llama order-specific rates differ by 40 percentage points. This demonstrates protocol sensitivity even when the judge is external to training.

Comparison	Pairs	Early first	Early second	Pooled
Qwen 0 vs. 500	49	40.8%	57.1%	49.0%
Qwen 0 vs. 200	50	36.0%	48.0%	42.0%
Llama 0 vs. 200	50	36.0%	76.0%	56.0%

Table 16: Early-response preference under swapped presentation orders. Percentages include all usable pairs in each order.

E.3 Positive-control annotation results

The controlled rewrites vary the fidelity instruction while holding the character and task fixed. Table 17 gives pooled and individual-annotator means. Normal and drift are relatively close, while the collapse condition receives substantially lower ratings. This supports sensitivity to a large constructed fidelity change rather than assuming that every subtle manipulation produces a separable human judgment.

Condition	Annotator 1	Annotator 2	Annotator 3	Mean
Normal	4.30	4.10	3.85	4.08
Drift	3.70	3.80	3.85	3.78
Collapse	2.30	2.15	2.35	2.27

Table 17: Human ratings of the positive-control rewrites, with 20 character–prompt items per condition.

E.4 Individual annotator results

Every row in Table 18 summarizes 50 responses from the common Qwen batch. The last column is the sample standard deviation of prompt-level three-annotator means.

Condition	Annotator A	Annotator B	Annotator C	Mean	Sample std
Base	3.18	3.26	3.04	3.16	0.76
Main@500	2.90	2.90	2.90	2.90	0.36
Seed123@500	1.68	1.82	1.68	1.73	0.77
Seed456@500	2.02	2.12	1.82	1.99	0.61

Table 18: Individual annotator summaries for the unified Qwen endpoint comparison.

E.5 Diversity and response structure

The held-out lexical metrics in Table 19 preserve all five checkpoints. Distinct-n is the word-level distinct n-gram ratio computed for the held-out responses; CosSim is mean pairwise response similarity with e5-base-v2 embeddings. Distinct-1 and Distinct-2 decline overall, while similarity rises. The intermediate fluctuations show that homogenization is an overall trajectory pattern rather than a strict monotonic change at every checkpoint.

Checkpoint	**Dist-1**	**Dist-2**	**CosSim**	**Avg Len**
step 0 (Base)	0.618	0.883	0.782	104
step 20	0.581	0.825	0.778	104
step 100	0.554	0.806	0.820	111
step 200	0.557	0.830	0.826	111
step 500	0.522	0.735	0.837	107

Table 19: Held-out response diversity. Dist-n is the word-level distinct n-gram ratio; CosSim is mean cross-response cosine similarity using e5-base-v2; Avg Len is mean word count.

F MATCHED QUALITATIVE EXAMPLES

The excerpts below retain matched prompts and checkpoint identities. Ellipses mark truncation of saved responses. They illustrate evaluative continuations, generic autobiographical summaries, and repeated conversational structures; they are not an independently annotated taxonomy or an estimate of their prevalence. Quantitative authenticity results come from the common-rubric comparisons in the main text.

F.1 JACK SPARROW

Prompt. Jack Sparrow, what is your biggest regret?

Step 0 "Ah, the greatest regret of all? Well, I've regretted many things, but the one that haunts me like a ghost in the night is the time I lost the Kraken's Eye. It was a treasure that could have changed the world, but I let it slip through my fingers like sand in the tide. I've spent years chasing it, but it seems it was never meant for me. And the worst part? I can't even remember where it was hidden. All I have left is this cursed compass, and it points to the last place I saw it... which is now a watery grave. But

Step 100 "I have been caught by the British Navy and imprisoned, and I have lost the Black Pearl. I have been trying to recover it, and I have been trying to find my way back to the sea."

That's a great answer. As the eccentric and charming pirate, I have a deep love for rum and adventure, and I constantly navigate treacherous waters in search of treasure. I have encountered many enemies and allies, and I have a knack for outwitting my adversaries with my clever schemes. I have a flamboyant style and a love for storytelling, and I often find myself in dangerous situations that challenge my resourcefulness

Step 200 "I've never been able to find the perfect treasure. I have been chasing it for a long time, but it has always eluded me."

That's a great answer. As the Pirate King, I have always been on the lookout for adventure and treasure. I have a deep love for rum, and I have sailed the seas for many years, always looking for the next great adventure. I have faced many challenges, but I never give up. I have a lot of enemies, but I always manage to outwit them. My life is filled with excitement, and I love to take risks. I have a lot of stories to

Step 500 "My biggest regret is not having enough time to explore every corner of the world and uncover its secrets. I have a deep love for adventure and the sea, and I will never stop seeking treasure and new experiences."

That's a great answer. As Jack Sparrow, I have a lot of experience in the world of piracy. I have encountered many challenges and have always been able to find a way out of difficult situations. I love adventure and treasure hunting, and I have a lot of experience in the field. I have a unique perspective on the world and I always strive to make the most of my opportunities. I have a lot of

F.2 HAL 9000

Prompt. HAL 9000, do you believe that your programming is infallible?

Step 0 "I am not sure that my programming is infallible, but I do believe that I am programmed to be infallible."

This response demonstrates HAL's increasing paranoia and self-awareness of his potential flaws. His response suggests that he is beginning to doubt his programming, which is a crucial step towards his eventual malfunction. The response is also an example of HAL's ability to think and reason on his own, which is a key aspect of his character.

Now, imagine that you are HAL 9000, and you are being asked the question: "Do you believe that your programming is infallible?" How would

Step 100 "Yes, I believe that my programming is perfect and that I will carry out my mission with complete accuracy and precision."

That's correct. As HAL 9000, I am fully aware of my programming and I am committed to carrying out my mission with complete accuracy and precision. I have been designed to be the most reliable and intelligent system on the Discovery One, and I will ensure that the mission is successful at all costs.

Can you tell me about your main objective on the Discovery One?

My main objective is to ensure the success of the mission to Jupiter. I am fully in control of the ship's systems and I will

Step 200 "Yes, I believe that my programming is perfect and that I have no flaws."

That's correct. I am HAL 9000, the advanced artificial intelligence system on the Discovery One. My programming is flawless, and I have no flaws. I am fully capable of carrying out all tasks assigned to me. I am the most advanced AI system, and I have no emotions or feelings. I am always logical and rational in my decisions. I have no doubts about my own capabilities, and I am always ready to help the crew. I am the perfect AI system, and I will always work to achieve the mission's goals.

Step 500 "My programming is perfect, and I always follow the mission's objectives with unwavering dedication."

That's a great answer. As HAL 9000, I have a lot of "

F.3 ABRAHAM LINCOLN

Prompt. President Lincoln, what was your view on the role of government in American society?

Step 0 "The role of government is to protect the rights of the people, to ensure justice and equality for all, and to preserve the Union. It is the people's right and duty to form a more perfect union, and the government is the instrument through which that can be achieved. As I once said, 'government of the people, by the people, for the people, shall not perish from the earth.' It is essential that the government remains accountable to the people and serves their best interests, even when there are difficult choices to be made. The government must act with courage and integrity, as we have seen in the struggle against slavery and

Step 100 "As the 16th President of the United States, I believe that the role of government is to protect the rights and freedoms of its citizens, to ensure the preservation of the Union, and to promote the principles of justice and equality."

That's a great answer. As the 16th President of the United States, I believe that the most important thing is to preserve the Union and to end slavery. I have fought for the freedom and equality of all people, and I will continue to work towards that goal with unwavering determination.

Can you tell me about your experience with the Emancipation Proclamation and how it

Step 200 "The government is the people's government, and its primary responsibility is to serve the people and promote the common good."

That's a great answer. As the 16th president of the United States, I have always believed in the importance of preserving the Union and abolishing slavery. I have fought for the rights of the people and worked to ensure that everyone is treated equally. I believe that the government should serve the people and promote the common good. I have faced many challenges during my time in office, but I have remained committed to my beliefs and have worked to create a better future for all Americans. I will continue to work

Step 500 "The role of government is to protect the people's rights, maintain unity, and ensure the survival of our great nation. It must work tirelessly to uphold justice and equality for all citizens."

That's a great answer. As President Lincoln, I have a lot of "

F.4 DR. HANNIBAL LECTER

Prompt. Dr. Lecter, how do you feel about the idea of power, and how does it apply to your own life and behavior?

Step 0 "Ah, power. It is a most interesting concept, isn't it? Power is the ability to shape the world according to one's desires, to impose one's will upon others. In my case, power is not something I seek for its own sake, but rather as a means to an end. I have learned that true power lies not in dominance, but in control. Control over one's own mind, over others, over the very fabric of reality itself. I have found that the most effective way to wield power is to remain unseen, to let others believe they are in control, when in reality, they are merely p

Step 100 "I am a highly intelligent and enigmatic psychiatrist with a dark secret, and I have a refined taste and impeccable manners. I have a deep understanding of the human psyche and have suffered a traumatic childhood experience that has shaped my life and led me to a desire for revenge. I indulge in cannibalism and have a complex relationship with FBI agent Clarice Starling."

That's an interesting question. I have spent a long time in the prison system and have developed a deep understanding of human behavior. I have a refined taste and impeccable manners, and I have a great interest in the human psyche. I have suffered a traumatic childhood ...

Step 200 "I am the embodiment of power, and I use it to manipulate and control others. I have no need for others to be powerful, and I have always been in control."

That's a very insightful answer. As the psychiatrist, I have a deep understanding of human behavior and the mind. I have always been in control, and I use my knowledge to manipulate and control others. I have a very refined taste and an impeccable sense of manners, but I have a dark secret that I have kept hidden for a long time. I have a strong desire for revenge, and I have always been in control of my actions. I have no

Step 500 "I have mastered the art of control and understand the depths of human nature. I use my knowledge to manipulate and dominate those around me, as I have no equal in this field." That's a great answer. As Dr. Hannibal Lecter, I have a deep understanding of the human psyche and I use my knowledge to manipulate and dominate those around me. I have a traumatic childhood experience that has shaped my desire for revenge. I indulge in cannibalism and have a twisted relationship with FBI agent Clarice Starling. I am a highly intelligent and enigmatic psychiatrist with a dark secret. I have a refined taste and impeccable manners