

# Fine-Tuning for Social-Media User Simulation: A Study Across Models and Evaluators

Sneheel Sarangi<sup>1,2 \*</sup>, Maximilian Puelma Touzel<sup>2</sup>, Aurélien Bück-Kaeffer<sup>1,2,3</sup>,  
Zachary Yang<sup>1,2,3</sup>, Jean-François Godbout<sup>2,4</sup>, Reihaneh Rabbany<sup>1,2</sup>

<sup>1</sup>McGill University    <sup>2</sup>Mila – Quebec Artificial Intelligence Institute

<sup>3</sup>Ubisoft La Forge    <sup>4</sup>Université de Montréal

## Abstract

Does fine-tuning on a user’s social-media history make an LLM a better simulator of that user? We study this question using QLoRA, conditioning models on a user’s chronological action history and a reply thread before generating that user’s reply. Across 12 matched baseline–QLoRA pairs from four model families (0.8B–32B), plus a baseline-only 122B model, LLM-as-judge believability evaluations show that QLoRA improves several smaller models, has mixed effects on stronger models, and does not lift a sub-10B model into the strongest large-model tier. However, a real-versus-generated classifier gives a different result: QLoRA makes Qwen, OLMo, and Llama replies much less detectable, while Gemma remains readily detectable. Contextual believability and detector AUC are uncorrelated across systems. We further explore LLM evaluations: strong LLM judges agree on broad quality tiers, and a blind 200-comparison audit finds human–judge agreement of 66% when three strong judges agree and 47% when they split. The measured effect of QLoRA therefore depends on the evaluation axis.

## 1 Introduction

Realistic social-media user simulation requires more than fluent or human-like text: a reply must be plausible for a particular user in the specific thread they are entering. This capability supports computational social science and controlled studies of social computing systems (Park et al., 2022; Gao et al., 2023; Park et al., 2023; Argyle et al., 2023; Sarangi et al., 2026). A useful user simulator should therefore condition on both the user’s behavioral history and the local conversational context.

Prior work often conditions LLMs on persona descriptions, profiles, or behavioral summaries (Li et al., 2024; Lei et al., 2026). This is flexible, but

it shifts much of the problem to profile construction and context selection, and persona-conditioned models can drift over longer interactions (Abdulhai et al., 2025). We study a complementary question: if we fine-tune directly on chronological social-media action histories, do models become better user simulators?

We run a controlled, baseline-matched study of QLoRA, a parameter-efficient fine-tuning method (Hu et al., 2021; Detrmers et al., 2023), with a recipe similar to that used in related social-media fine-tuning work (Bück-Kaeffer et al., 2025; Pagan et al., 2025). Each example contains a user’s chronological posts, replies, likes, and reposts, followed by a target thread and the user’s observed reply. This formulation tests whether models can learn directly from heterogeneous behavioral traces without a separately constructed persona summary. Holding the history, thread, and generation setup fixed within each matched pair isolates the effect of parameter adaptation across four model families, with QLoRA comparisons from 0.8B to 32B parameters and an additional baseline-only 122B model.

We find that QLoRA improves several smaller models but has mixed or negative effects on stronger baselines, and no tuned model below 10B reaches the strongest large-model tier. The conclusion also depends on how simulation quality is measured: LLM judges favor models that produce contextually believable replies, whereas a real-versus-generated classifier shows that QLoRA makes replies from Qwen, OLMo, and Llama substantially less detectable even when contextual believability does not improve. These two evaluation axes are uncorrelated across systems. Human audits further show that agreement among strong LLM judges is informative: human–judge agreement is substantially higher when the judges agree than when they split.

Our contribution is therefore not only a con-

\*Corresponding author: sneheelsarangi@gmail.com

trolled comparison of QLoRA across model families and scales, but also an evaluation of what different notions of “better user simulation” capture. By holding user-specific context fixed within matched comparisons and jointly examining contextual believability, distributional discriminability, judge agreement, and human judgments, we separate improvements in surface imitation from improvements in contextual fidelity. The results show that adaptation, model capacity, and evaluator choice can lead to different conclusions about simulation quality.

## 2 Related Work

User simulation has roots in dialogue systems (SCHATZMANN et al., 2006) and now includes LLM-based social and economic agents (Bück-Kaeffer et al., 2025; Horton et al., 2023; Park et al., 2023; Argyle et al., 2023). Social-media user modeling is closest to our setting: FineRob studies behavior simulation from social traces (Li et al., 2024), while HumanLLM and HumanLM target broader individualized behavior modeling (Lei et al., 2026; Wu et al., 2026). Related personalization work studies how much history should be placed in context or compressed into parameters (Salemi et al., 2024; Gu, 2026; Zhang and Wang, 2026; Zhang et al., 2026). Our contribution is a controlled analysis of one standard QLoRA recipe across model families and scales, using raw multi-action histories and the original reply thread. Every model and evaluator receives the same user-specific context within a comparison, which separates the effect of parameter updates from the effect of supplying more informative prompts.

Exact-reference matching is too narrow for open-ended replies. PersonaGym (Samuel et al., 2025), computational Turing tests (Pagan et al., 2025), and Blueprint (Bück-Kaeffer et al., 2025) motivate evaluating both human-likeness and fidelity to reference behavior. LLM judges support context-conditioned comparison (Zheng et al., 2023; Liu et al., 2023) but exhibit evaluator biases (Wang et al., 2023). Text classifiers instead measure whether generated text remains distinguishable from real text (Dawkins et al., 2025). Prior pooled detectors classify isolated replies across authors; our pairwise judges assess whether a reply is plausible for a specified user and thread, while our per-system classifiers expose distributional artifacts in reply text. These are related questions with dif-

ferent failure modes. We compare them directly, randomize candidate order, and analyze judge and human agreement. Appendix A discusses further related work.

## 3 Method

**Task.** Given a user’s chronological action history  $H$  and a thread stimulus  $S$ , the model generates a reply  $\hat{r}$  intended to match the user’s actual reply  $r^*$ . The system message contains the user’s action history; the user message contains the thread and the suffix “The user replied to the above post with:”; the assistant message is the gold reply. Training loss is applied only to the Assistant response.

**Data.** We use Bluesky action histories (Kleppmann et al., 2024) from the data source described by Bück-Kaeffer et al. (2025), collected around March 2025 from Canadian and American users active in political discussions. Users are retained if they have at least 20 context events and 10 authored post/reply events. Histories include posts, replies, likes, and reposts, ordered chronologically and capped by token budget. The processed collection contains 8,616 reply examples from 1,117 users. For training, we use a fixed 5,000-example subset, and validation uses 1,000 examples. Splits combine held-out users and temporal holdout, so evaluation covers both unseen users and future behavior from seen users. The main comparison uses 2048-token histories. Appendix M evaluates shorter and longer histories. Appendix E repeats the comparison on random users from a non-political Bluesky corpus (Failla and Rossetti, 2024, 2025).

**Models and training.** We evaluate instruction-tuned Qwen3 and Qwen3.5 (Yang et al., 2025; Qwen Team, 2026), Gemma4 (Gemma Team, 2026), OLMo3 (Team Olmo et al., 2025), and Llama3.1 (Grattafiori et al., 2024) models from 0.8B to 122B parameters. Scale plots use the loaded checkpoint parameter counts: Gemma4-E2B and Gemma4-E4B contain approximately 5.2B and 8B total parameters. The official Qwen checkpoint is Qwen3.5-122B-A10B; figures and tables abbreviate it as Qwen3.5-122B-A10. All fine-tuning used QLoRA (Hu et al., 2021; Dettmers et al., 2023) with rank 32,  $\alpha = 64$ , no dropout, and an effective batch size of 8. Qwen3.5-122B-A10 is baseline-only.

**Contextual evaluation.** We evaluate 1,276 test examples per system. Qwen3.5-27B-FP8 scores

content, stance, style, and believability on 1–5 scales. We also report TwHIN-BERT cosine similarity (Zhang et al., 2023), lexical metrics, and length ratio. In pairwise tournaments, each judge sees the same history and thread with two randomly ordered replies. The main ELO averages separate 14,300-comparison tournaments run by Qwen3.5-122B-A10, Gemma4-31B, and Qwen3.5-27B; judge-specific rankings are provided in the Appendix G for further analysis.

**Discriminability.** For each system, a balanced real-versus-generated classifier combines frozen mean-pooled TwHIN-BERT embeddings with length, punctuation, URL, hashtag, mention, emoji, and formatting features, similar to the classifier used by Pagan et al. (2025). Users are separated across train and test. ROC-AUC is primary: 0.5 is chance, and lower is harder to discriminate from human responses. For each test context, we compare the classifier’s realness scores for two candidate replies and select the higher-scoring reply. Applying the same ELO update used for judge rankings produces a detector ladder without further annotations.

**Human audit.** Two authors independently rate the same 200 comparisons; six non-author raters each rate a shared 30-comparison validation set. Raters see the user history, thread, and randomized candidates, and choose a preference. The comparisons span clear and contested model pairs as well as real-generated pairs. The audit question operationalizes whether a response is likely to have been written by the specified user in that thread. Appendices B, G, H, and I give full details.

## 4 Results

**QLoRA does not improve performance uniformly.** Figure 1 shows positive scalar-believability changes for 6 of 12 matched systems, including a negligible +0.03 for Gemma4-E4B; the other 6 degrade or do not change. Clear gains occur for Qwen3.5-0.8B, Qwen3.5-2B, Qwen3.5-4B, OLMo3-7B, and Llama3.1-8B. Gemma4-E2B degrades, Qwen3.5-27B loses 0.69 points (4.74 to 4.05), OLMo3.1-32B loses 0.08, and Gemma4-31B stays near its baseline. ELO and the other scalar metrics show related but model-specific changes. Adaptation helps when a baseline has headroom but can reduce contextual fit for capable models. Full fine-tuning checks on

| Model        | $\Delta$ Bel. | $\Delta$ Judge ELO | $\Delta$ AUC |
|--------------|---------------|--------------------|--------------|
| Qwen3.5-0.8B | +0.27         | +112               | −.256        |
| Qwen3.5-2B   | +0.79         | +313               | −.328        |
| Qwen3.5-4B   | +0.39         | +148               | −.276        |
| Qwen3.5-9B   | −0.03         | −35                | −.257        |
| Qwen3.5-27B  | −0.69         | −138               | −.279        |
| Gemma4-E2B   | −0.66         | −218               | +0.059       |
| Gemma4-E4B   | +0.03         | +100               | +0.014       |
| Gemma4-31B   | .00           | +43                | +0.057       |
| OLMo3-7B     | +0.34         | +156               | −.264        |
| OLMo3.1-32B  | −0.08         | −17                | −.289        |
| Llama3.1-8B  | +0.30         | +86                | −.270        |

Table 1: Matched change from base to QLoRA across evaluation axes. Positive believability and judge-ELO changes favor QLoRA; negative AUC changes mean that generated replies become harder to distinguish from real replies.

4B and 7B models provide little additional benefit (Appendix K).

### QLoRA changes the evaluation axes differently.

Table 1 places the matched intervention effects side by side. Believability and judge ELO improve for some systems and decline for others. Detector AUC follows a more regular family pattern: it falls for all eight Qwen, OLMo, and Llama pairs with complete results, but rises for all three Gemma pairs. QLoRA’s most consistent effect is therefore a family-dependent shift toward the surface form of human replies, not a uniform improvement in contextual simulation.

### Larger base models outperform smaller tuned models.

Under the strong-judge aggregate ELO, the top four systems are Gemma4-31B LoRA (1867), Gemma4-31B base (1825), Qwen3.5-122B-A10 base (1783), and Qwen3.5-27B base (1764). No tuned model below 10B reaches this tier. Llama3.1-8B LoRA, the best sub-10B tuned system (1588), remains over 170 points below these large baselines. Scalar metrics agree: Qwen3.5-27B base reaches 4.74 believability, 2.48 content, and 2.71 style, against 3.74, 2.13, and 2.63 for Llama3.1-8B LoRA. These gaps exceed the useful within-model QLoRA gains. Full results and uncertainty appear in Appendix Tables 15 and 14.

### Evaluation agreement scales with judge size.

Figure 2 shows that pairwise judge agreement increases across model-size tiers. Kendall  $\tau$  between a judge’s system ranking and the main three-judge aggregate rises from 0.51–0.57 for sub-4B judges to 0.66–0.80 for the 27B–122B tier; agreement between pairwise win rates rises from 0.82 to 0.95. GPT-5.4 and GPT-5.4-mini, two API judges from a

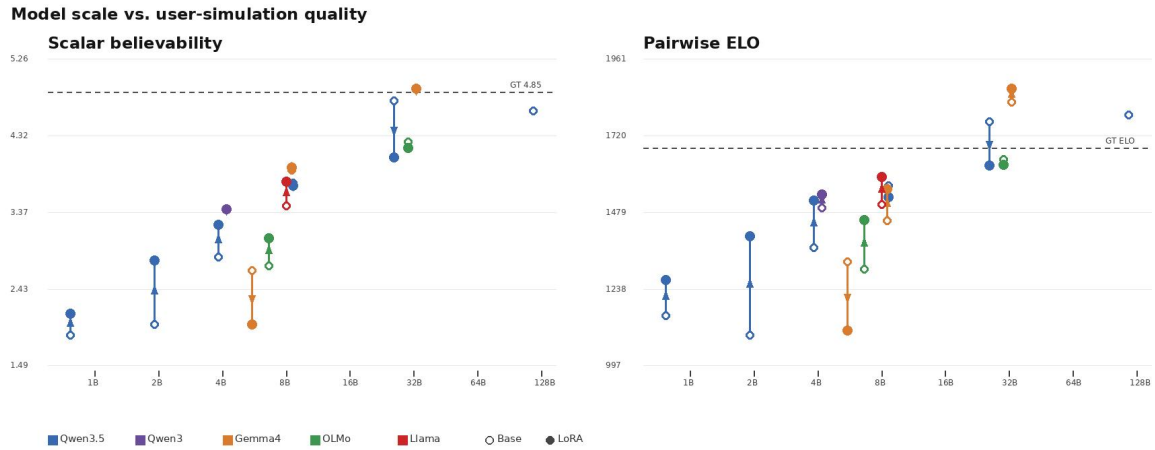


Figure 1: Model scale versus scalar believability and pairwise ELO. The x-axis is log-scaled. Open circles are base models; filled circles are QLoRA adapters; arrows connect matched base–LoRA systems and point toward the LoRA result. Dashed lines mark ground-truth scalar believability and aggregate ELO.

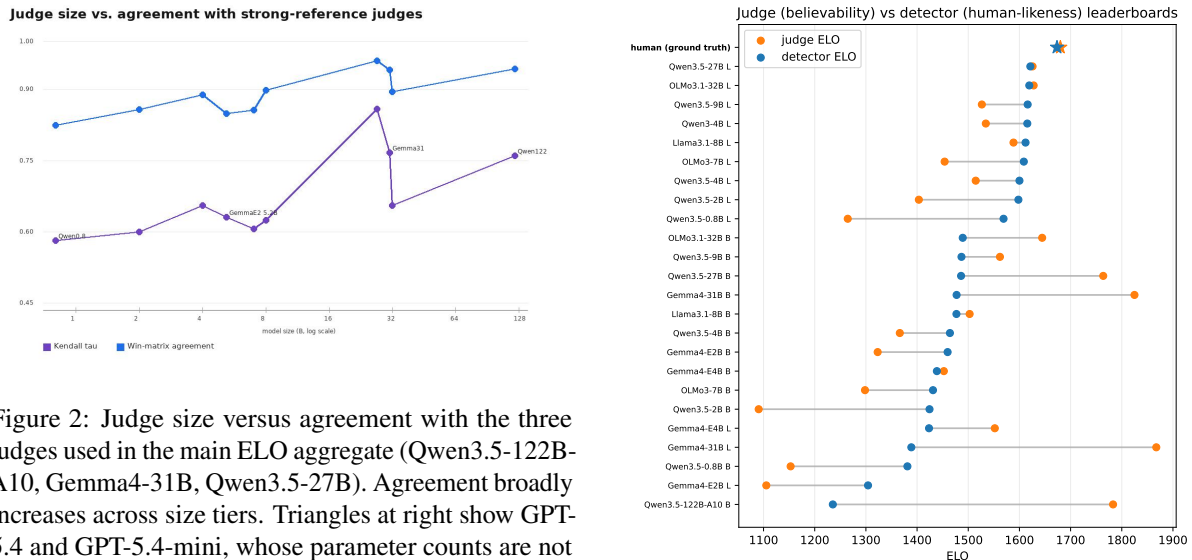


Figure 2: Judge size versus agreement with the three judges used in the main ELO aggregate (Qwen3.5-122B-A10, Gemma4-31B, Qwen3.5-27B). Agreement broadly increases across size tiers. Triangles at right show GPT-5.4 and GPT-5.4-mini, whose parameter counts are not public.

family absent from the candidate set, reach  $\tau = .82$  and  $.75$  (OpenAI, 2026). The pattern is therefore not confined to same-family Qwen judges. Weaker judges also show larger self-preference bias: Gemma4-E2B raises its own base-plus-LoRA win rate by 0.20 relative to the main three-judge assessment, while the three strong judges show lifts below 0.08 (Appendix Table 6). Broad quality tiers are stable under capable judges, but middle ranks and ground-truth placement remain evaluator-dependent. Appendix Figure 5 gives the full rankings.

**Discriminability gives a different ordering.** In general, base-model replies are readily distinguishable from human replies (AUC 0.90–0.98).

Figure 3: The contextual LLM-judge and text-discriminability ELO ladders differ substantially. B and L denote base and QLoRA systems. Full ranks are in Appendix Table 8.

QLoRA reduces AUC to 0.64–0.73 for Qwen, OLMo, and Llama, while Gemma remains at 0.94–0.98. Length matching preserves the effect: Qwen3.5-27B changes from .885 to .629, close to its raw .921-to-.643 reduction. Judge believability and detector AUC are uncorrelated across systems (Spearman  $\rho = .02$ ,  $p = .93$ , bootstrap 95% CI  $[-.44, .48]$ ). Their ELO ladders also differ (Figure 3): human replies rank first under the detector, followed by Qwen3.5-27B LoRA and OLMo3.1-32B LoRA, while the judge-leading

Gemma4-31B LoRA ranks 22nd. A classifier augmented with history- and thread-similarity features still selects the real reply 62–70% of the time for Qwen, OLMo, and Llama adapters and 93–97% for Gemma. QLoRA can improve surface imitation without making a reply indistinguishable from the user’s real response in context. Appendix H includes further details.

**Qualitative Analysis.** We inspected 44 Qwen3.5-27B base-versus-QLoRA comparisons using GPT-5.4 rationales; the base wins 27 and QLoRA wins 17. QLoRA losses most often involve unsupported specifics (7 cases), fragments or awkward coherence (5), off-thread drift (5), and generic or polished language (4). Base losses most often involve generic, polished, or overlong advocacy (5). QLoRA captures terse register, which helps detection, but also reproduces unhedged details and loose thread logic. Appendix J discusses this further.

**Human audit.** The clearest evidence of ambiguity comes from 100 real-generated comparisons selected because Qwen3.5-122B-A10 preferred the generated reply. Across the two authors’ 200 votes on these cases, 41.5% favored the generated reply, 29.5% favored the real reply, and 29.0% were ties. The judge’s choice was therefore not a clear human error, but it was not a clear human preference either.

The broader audit shows that reliability varies with comparison difficulty. The authors agree on 90% of high-ELO difference model pairs ( $\kappa = .80$ ), compared with 55–67% across the remaining stratified slices ( $\kappa = .27-.33$ ) and 42% on the adversarial real-generated slice ( $\kappa = .12$ ). Individual human-judge agreement is 66% when the three strong judges are unanimous and 47% when they split; the corresponding human-majority rates are 81% and 55%. Humans also select the higher-ELO system more often as the gap grows: 45% below 150 ELO points, 60% at 150–300, and 68% above 300. Appendix I reports the design, denominators, external-rater comparison, and full breakdown.

## 5 Discussion

The contextual metrics establish a practical hierarchy. Across LLM-as-judge ratings, large baselines and large fine-tuned models lead, and QLoRA does not let sub-10B models match them. Its gains for smaller models make it an efficiency tool within

a capacity tier; base-model capacity remains decisive. The non-political replication in Appendix E supports this pattern, including the case that the 27B model QLoRA showed worse performance than the baseline.

At the same time, QLoRA reduces detectability for Qwen, OLMo, and Llama, including the 27B model that loses 0.69 believability points. Gemma remains highly detectable even when its 31B adapter leads the judge ranking. The conditioned classifier and qualitative provide insights: adaptation captures terse register while leaving context-fit errors such as unsupported specifics, fragments, and off-thread drift. Surface imitation can improve while contextual coherence declines.

The human audit separates hard cases from general evaluator failure. On judge-selected generated-over-real cases, votes divide among generated, real, and tie, indicating that the observed ground-truth losses are genuinely ambiguous. Across the broader set, humans increasingly favor higher-ELO systems as the gap grows, and unanimous judge decisions align better with humans than split decisions. Three-judge consensus is therefore a useful confidence signal, although it is not infallible.

Evaluations of contextual user simulation should use capable judges from different families, report their agreement, and audit strata that include judge disagreements. A separate detector should report discriminability with real-versus-real, length-matched, and context-conditioned controls. The two scores should remain separate because they measure different properties. Under a compute constraint, QLoRA remains useful for improving smaller models; strong contextual simulation still requires sufficient base-model capacity.

## 6 Conclusion

QLoRA improves many smaller models but does not close the gap to strong large baselines and can harm capable systems. Its effect also depends on the evaluation axis: for Qwen, OLMo, and Llama, adaptation can make replies harder to distinguish from human text without improving contextual believability, with Gemma being an outlier. These results separate surface imitation from contextual user simulation. Strong LLM judges agree on broad quality tiers, and human agreement is higher when those judges agree than when they split. User-simulation evaluations should therefore report both evaluation axes and judge agreement.

## Limitations

The main corpus is English-language Bluesky data from North American political discussions. The non-political replication tests content-domain transfer, but cross-platform and cross-lingual validity remain open. The cross-model comparison holds the QLoRA configuration fixed to isolate model-family and scale effects; adaptation outcomes may differ under other configurations or training methods. The classifier captures detectability from reply text and may reward bland or repetitive outputs. Context-conditioned controls reduce this concern but still select the real reply above chance. Finally, the full 200-comparison audit was rated by two paper authors; the six non-author raters evaluated a shared 30-comparison validation set. The audit supports agreement-based confidence estimates and does not define a definitive human leaderboard.

## Ethics Statement

This work studies public social-media behavior for user simulation, which creates privacy and misuse risks if models are used to impersonate or target individuals. The source posts were publicly accessible through the AT Protocol, which is designed for public conversation and third-party applications. Bluesky’s [Terms of Service](#) state that users retain ownership of their content and that Bluesky does not control third-party use. We therefore minimize downstream exposure: identifiers are anonymized during training following [Bück-Kaeffer et al. \(2025\)](#), private content is excluded, and we do not release post text, fine-tuning data, or trained weights. Researchers reusing the collection should independently verify the platform terms in force at the time of access. Our results also show that capable untuned models can already produce plausible personalized replies, so withholding adapters does not remove the broader impersonation risk.

## Acknowledgments

The authors acknowledge funding support from the Future of Life Institute, IVADO, and the Canada CIFAR AI Chairs Program. This work benefited from awards from the Fonds de recherche du Québec held by Sneheel Sarangi (<https://doi.org/10.69777/2006614>) and Aurélien Bück-Kaeffer (<https://doi.org/10.69777/2004934>). The authors sincerely thank the anonymous reviewers for their constructive feedback and for helping

improve the paper. We also thank the members of the Complex Data Lab for helpful discussions.

## References

- Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. [Consistently simulating human personas with multi-turn reinforcement learning](#). *Preprint*, arXiv:2511.00222.
- Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.
- Aurélien Bück-Kaeffer, Je Qin Chooi, Dan Zhao, Maximilian Puelma Touzel, Kellin Pelrine, Jean-François Godbout, Reihaneh Rabbany, and Zachary Yang. 2025. [BluePrint: A social media user dataset for llm persona evaluation and training](#). *Preprint*, arXiv:2510.02343.
- Hillary Dawkins, Kathleen C. Fraser, and Svetlana Kiritchenko. 2025. [When detection fails: The power of fine-tuned models to generate human-like social media text](#). In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *Preprint*, arXiv:2305.14314.
- Andrea Failla and Giulio Rossetti. 2024. [“i’m in the bluesky tonight”: Insights from a year worth of social data](#). *PLOS ONE*, 19(11):e0310330.
- Andrea Failla and Giulio Rossetti. 2025. [Bluesky social dataset](#).
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. [S<sup>3</sup>: Social-network simulation system with large language model-empowered agents](#). *Preprint*, arXiv:2307.14984.
- Gemma Team. 2026. [Gemma 4 technical report](#). *arXiv preprint arXiv:2607.02770*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.

- Shangding Gu. 2026. [Long context, less focus: A scaling gap in llms revealed through privacy and personalization](#). *Preprint*, arXiv:2602.15028.
- Daniel Han and Michael Han. 2024. Unsloth: Fast LoRA fine-tuning framework. <https://github.com/unslothai/unsloth>.
- John J. Horton, Apostolos Filippas, and Benjamin S. Manning. 2023. [Large language models as simulated economic agents: What can we learn from homo silicus?](#) *Preprint*, arXiv:2301.07543.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Martin Kleppmann, Paul Frazee, Jake Gold, Jay Graber, Daniel Holmgren, Devin Ivy, Jeromy Johnson, Bryan Newbold, and Jaz Volpert. 2024. [Bluesky and the at protocol: Usable decentralized social media](#). In *Proceedings of the ACM Conext-2024 Workshop on the Decentralization of the Internet*, CoNEXT '24, pages 1–7. ACM.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Yuxuan Lei, Tianfu Wang, Jianxun Lian, Zhengyu Hu, Defu Lian, and Xing Xie. 2026. [Humanllm: Towards personalized understanding and simulation of human nature](#). *Preprint*, arXiv:2601.15793.
- Kun Li, Chenwei Dai, Wei Zhou, and Songlin Hu. 2024. [Fine-grained behavior simulation with role-playing large language model on social media](#). *Preprint*, arXiv:2412.03148.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.
- Bo Ni, Branislav Kveton, Samyadeep Basu, Subhojyoti Mukherjee, Leyao Wang, Franck Dernoncourt, Sungchul Kim, Seunghyun Yoon, Zichao Wang, Ruiyi Zhang, Puneet Mathur, Jihyung Kil, Jiuxiang Gu, Nedim Lipka, Yu Wang, Ryan A. Rossi, and Tyler Derr. 2026. [Reasoning-based personalized generation for users with sparse data](#). *Preprint*, arXiv:2602.21219.
- OpenAI. 2026. [Gpt-5.4 model documentation](#). OpenAI API model documentation.
- Nicolò Pagan, Petter Törnberg, Christopher A. Bail, Anikó Hannák, and Christopher Barrie. 2025. [Computational turing test reveals systematic differences between human and ai language](#). *Preprint*, arXiv:2511.04195.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). *Preprint*, arXiv:2304.03442.
- Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. [Social simulacra: Creating populated prototypes for social computing systems](#). *Preprint*, arXiv:2208.04024.
- Qwen Team. 2026. [Qwen3.5](#). Hugging Face model card.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. [LaMP: When large language models meet personalization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392. Association for Computational Linguistics.
- Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. [PersonaGym: Evaluating persona agents and LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022. Association for Computational Linguistics.
- Sneheel Sarangi, Maximilian Puelma Touzel, Aurélien Bück-Kaeffer, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. 2026. [Ease configuration facilitates a reproducible science of llm social simulations](#). *Preprint*, arXiv:2605.30258.
- JOST SCHATZMANN, KARL WEILHAMMER, MATT STUTTLE, and STEVE YOUNG. 2006. [A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies](#). *The Knowledge Engineering Review*, 21(2):97–126.
- Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. [Olmo 3](#). *Preprint*, arXiv:2512.13961.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. [Large language models are not fair evaluators](#). *Preprint*, arXiv:2305.17926.

Shirley Wu, Evelyn Choi, Arpandeeep Khatua, Zhanhan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. 2026. [Humanlm: Simulating users with state alignment beats response imitation](#). *Preprint*, arXiv:2603.03303.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

Chenheng Zhang, Yijun Lu, Lizhe Fang, Chunyuan Zheng, Jiajun Chai, Xiaohan Wang, Guojun Yin, Wei Lin, Yisen Wang, and Zhouchen Lin. 2026. [Rethinking personalization in large language models at the token level](#). *Preprint*, arXiv:2603.06595.

Xinliang Frederick Zhang and Lu Wang. 2026. [Tsubasa: Improving long-horizon personalization via evolving memory and self-learning with context distillation](#). *Preprint*, arXiv:2604.07894.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. [TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations at twitter](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5597–5607. ACM.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

We used generative AI tools, including OpenAI ChatGPT, for language editing and to assist with figure generation from experiment outputs.

## A Extended Related Work

### A.1 User Simulation with Language Models

User simulation has a long history in dialogue systems research (SCHATZMANN et al., 2006). With instruction-tuned LLMs, the question has expanded from task-oriented dialogue to broader simulations of people, communities, and institutions. Aher et al. (2023) proposed Turing Experiments for testing whether LLMs can reproduce findings from human-subject studies, while Horton et al. (2023) studied LLMs as simulated economic agents. Park et al. (2023) introduced generative agents that simulate day-to-day human behavior using LLM-based architectures with memory retrieval. Argyle et al. (2023) showed that LLMs can reproduce survey

response distributions of demographic subgroups when prompted with appropriate conditioning information.

For social media specifically, Li et al. (2024) constructed the FineRob dataset of 1,866 users across three platforms and proposed observation-and-memory chain-of-thought fine-tuning to improve behavior prediction. Their work demonstrated that models that reason over past behaviors outperform those relying on static user profiles. Lei et al. (2026) introduced HumanLLM, a foundation model trained on diverse user modeling tasks including profile generation, social question answering, and writing imitation. Similarly, Wu et al. (2026) proposed reinforcement learning on latent psychological states to produce user simulators that are more consistent along dimensions like opinion and emotion.

We use raw chronological action histories as the conditioning signal, without curated persona descriptions or explicit psychological-state labels. This design tests whether standard QLoRA training can extract behavioral patterns directly from unstructured activity traces.

### A.2 Personalization and Context Length

Personalization work has examined how context length affects model quality. Salemi et al. (2024) introduced LaMP, a benchmark suite for personalized language modeling across retrieval and generation tasks. Gu (2026) introduced PAPERBench, a benchmark spanning 1K to 256K tokens of user background context, and found a consistent scaling gap: personalization quality degrades as context grows, even for models supporting long windows. They attribute this to limitations of soft attention under fixed capacity, showing that additional history does not guarantee better personalization.

Zhang and Wang (2026) addressed this through context distillation, compressing long user histories into model parameters via self-learning. Their approach separates the “knowing” (parametric memory) from “recalling” (episodic retrieval), reducing inference costs while maintaining personalization quality. Zhang et al. (2026) proposed token-level personalization signals that identify which output tokens depend on user-specific information and upweight them during training. Ni et al. (2026) explored context expansion through graph-based link prediction combined with reasoning alignment via fine-tuning.

We test a standard QLoRA recipe across model

families, scales, and amounts of behavioral context. The limited benefit from longer training histories is analogous to the “long context, less focus” phenomenon reported by Gu (2026), which concerns inference-time context length.

### A.3 Evaluation of Generated Personas

Evaluating persona fidelity requires multiple measures. Samuel et al. (2025) introduced PersonaGym and PersonaScore, an automatic evaluation framework covering linguistic habits, persona consistency, and action justification, validated against human judgments. Abdulhai et al. (2025) defined three consistency metrics (prompt-to-line, line-to-line, and Q&A consistency) and showed that LLM-generated personas drift from their assigned behavior over multi-turn interactions. Closest to our social-media setting, Pagan et al. (2025) proposed a computational Turing test for comparing AI and human language across platforms, and Bück-Kaeffer et al. (2025) introduced BluePrint for social-media persona evaluation and training. These works motivate separate measures of human-likeness and fidelity to reference behavior.

LLM judges (Zheng et al., 2023) support open-ended evaluation where reference-based metrics are insufficient. G-Eval showed that GPT-based evaluators can align better with human ratings than traditional NLG metrics in summarization and dialogue (Liu et al., 2023). Prior work also documents order and evaluator biases in pairwise judging (Wang et al., 2023). We randomize candidate order, report parsing reliability, compare multiple judges, and analyze cross-judge agreement.

## B Expanded Method and Evaluation Details

**Task formatting.** Each example is formatted as a three-turn chat. The system message contains the chronological user action history, the user message contains the reply thread and the suffix “The user replied to the above post with:”, and the assistant message contains the observed user reply.

**History construction.** For each target reply, we include only prior actions by the same user. Posts include authored text, replies include the parent post and the user’s response, and likes/reposts include the immediate parent post as compressed evidence of interest. Events are ordered from oldest to newest. When histories exceed the token budget,

we first remove older likes and reposts, then trim the oldest remaining events.

**Loss masking.** Training loss is applied only to assistant reply tokens. System and user tokens are conditioning context and receive the ignore label. For a sequence where the assistant reply begins after position  $h$ , the loss is  $-\frac{1}{n-h} \sum_{i=h+1}^n \log p_{\theta}(x_i | x_{<i})$ .

**Generation and scalar judging.** Scalar judging uses Qwen3.5-72B-FP8 to score content similarity, stance similarity, style similarity, and believability on 1–5 scales. Content, stance, and style compare the generated and gold replies; believability uses the history, thread, and generated reply.

**Reference metrics and baselines.** We compute TwHIN-BERT cosine similarity between generated and gold replies, plus length ratio, Levenshtein similarity, and Jaccard similarity. Each baseline receives the same action history and thread as its matched LoRA system, so baseline–LoRA comparisons isolate the effect of parameter updates while holding the input context fixed.

## C Training Details

**LoRA Configuration.** We use rank  $r = 32$ , scaling  $\alpha = 64$ , and no dropout. Target modules are q\_proj, k\_proj, v\_proj, o\_proj, gate\_proj, up\_proj, and down\_proj. Base models are loaded in 4-bit quantization via Unsloth (Han and Han, 2024) for QLoRA training.

**Optimizer.** QLoRA runs use AdamW with learning rate  $2 \times 10^{-4}$ , a linear schedule, 100 warmup steps, and weight decay 0.01. Cross-model results use each adapter’s epoch-10 checkpoint.

**Sequence lengths.** The main cross-model runs use 2,048-token histories and a maximum sequence length of 3,500 tokens, leaving room for the thread stimulus and reply. All main QLoRA runs use an effective batch size of 8. Training uses four H100 GPUs.

**Implementation.** Training uses multi-GPU data parallelism where needed for the larger models. Generation and judging use vLLM with tensor or data parallelism depending on model size.

### C.1 Evaluation

**Generation.** At inference time, we use vLLM (Kwon et al., 2023) with the same chat template

used during training to ensure prompt consistency. We evaluate every system on  $n = 1,276$  test examples. The main cross-model comparison uses 2048-token histories at evaluation time.

**LLM Judge Metrics.** We use Qwen3.5-27B-FP8 as an automated scalar judge, scoring predicted replies on four dimensions:

- **Content similarity** (1–5): topical overlap, entities, and facts referenced compared to the gold reply.
- **Stance similarity** (1–5): agreement, opinion, sentiment, and position toward the topic.
- **Style similarity** (1–5): tone, register, length patterns, and wording habits.
- **Believability** (1–5): whether the reply is a logical, in-character continuation given the user’s history and the thread context.

Content, stance, and style are evaluated from the predicted and gold reply texts. Believability uses the full context (history + thread + predicted reply) to assess coherence with the user’s established patterns.

**Embedding and Lexical Metrics.** We compute cosine similarity between mean-pooled embeddings of the predicted and gold replies using TwHIN-BERT (Zhang et al., 2023), a socially enriched language model trained on tweet representations. We also report length ratio (predicted / gold), Levenshtein similarity, and Jaccard similarity over word sets.

**Pairwise Believability ELO.** Scalar ratings can be sensitive to calibration differences, so we also run pairwise believability tournaments over pre-generated replies. For each comparison, the judge receives the action history, thread stimulus, and two candidate replies in randomized A/B order, then chooses the more believable continuation. We convert pairwise wins into ELO ratings with initial rating 1500 and  $K = 32$ . The main ELO column is a system-level aggregate: we run a complete ELO tournament separately for Qwen3.5-122B-A10, Gemma4-31B, and Qwen3.5-27B, then average each system’s resulting ELO across the three judges. We center the main ranking on this aggregate to reduce dependence on any single evaluator. We do not pool response-level comparisons across judges before fitting ELO. Appendix G reports

| Model       | Users  | $n$ | QLoRA win [95% CI] |
|-------------|--------|-----|--------------------|
| Qwen3.5-9B  | seen   | 499 | .643 [.602, .685]  |
|             | unseen | 192 | .594 [.530, .655]  |
| Qwen3.5-27B | seen   | 499 | .373 [.332, .414]  |
|             | unseen | 190 | .342 [.282, .404]  |
| Gemma4-31B  | seen   | 498 | .635 [.590, .675]  |
|             | unseen | 192 | .625 [.556, .695]  |

Table 2: Pairwise QLoRA win rate against its matched raw-history baseline on non-political Bluesky data. Intervals are user-clustered; all exclude .5.

judge-specific settings, rankings, and agreement across non-LoRA judges.

**Statistical uncertainty.** Appendix Table 14 reports 95% confidence intervals for scalar means over test examples. For aggregate ELO, it reports the range of system-level ratings assigned by the three strong judges.

**Baselines.** Each baseline uses the same instruction-tuned model without LoRA adapters, receiving the identical action history in the system message and thread in the user message. Thus, baseline-vs-LoRA comparisons isolate the effect of parameter updates while holding the input context fixed.

## D Data Processing Details

We use full-branch thread context, anonymized user IDs, no sibling replies, and token-capped user histories. Filtering requires at least 20 context events and at least 10 authored post/reply events per user. Splits combine user holdout and temporal holdout: 10% of users are fully held out, while examples for remaining users are ordered temporally and split into train/validation/test.

## E Non-Political Bluesky Replication

We repeated the matched base–QLoRA comparison on a non-political held-out-reply task constructed from random users in the Bluesky V3 corpus (Failla and Rossetti, 2024, 2025). The training subset contains approximately 6,000 examples and uses the main experiment’s QLoRA configuration, 2048-token histories, and Qwen3.5-27B-FP8 judge. We evaluate future replies from seen users and replies from users excluded from training. Table 2 reports the direct QLoRA-versus-base win rate, with confidence intervals obtained by resampling users.

The directions for Qwen3.5-27B and Gemma4-31B match the main experiment on both user splits.

Qwen3.5-9B shifts in favor of QLoRA, so the replication supports capacity dependence without implying a fixed threshold. We excluded IDs for which the 9B baseline entered repetition loops, applying the same exclusion to every model: 303 of 1,484 seen-user items and 51 of 245 unseen-user items. Because 93% of these failures came from the 9B baseline, this rule is conservative for its QLoRA comparison.

## F Pairwise Judge Prompt

## G Pairwise ELO Calculation and Judge Agreement

**Tournament construction.** The pairwise tournament uses 26 labels: 12 baseline systems, 12 matched QLoRA systems, Qwen3.5-122B-A10 as a baseline-only system, and the human ground-truth reply. This gives 325 unordered label pairs. We sample 44 test examples per pair, for 14,300 comparisons per judge before invalid-parse filtering. Candidate order is deterministically randomized with a fixed seed. Judges return structured JSON containing the winner (A or B) and a short rationale. Parse success exceeds 99.7% for every judge except Qwen3.5-0.8B, which parses 62.1% of comparisons; invalid responses are excluded from ELO fitting. All systems start at ELO 1500 and valid comparisons are processed in a deterministic shuffled order with  $K = 32$ .

**Aggregate ELO.** The main ranking averages system ratings from three complete tournaments judged by Qwen3.5-122B-A10, Gemma4-31B, and Qwen3.5-27B. We compute ELO independently for each judge before averaging system ratings. The appendix ELO range is the minimum and maximum across those three judges; it is not a bootstrap confidence interval. The following tables report judge-specific rankings, agreement, and diagnostics.

The judge-specific rankings form broad tiers even when exact ordering changes. Gemma4-31B LoRA, Qwen3.5-122B-A10, Qwen3.5-27B, and Gemma4-31B baseline are consistently upper-tier systems, with mean judge-specific ranks between 2.4 and 4.8. The weakest systems are also stable: Qwen3.5-0.8B baseline, Gemma4-E2B LoRA, and Qwen3.5-2B baseline have mean ranks above 23. Middle-tier systems vary more by judge; for example, Qwen3-4B baseline ranges from rank 1 under OLMo3-7B to rank 17 under Gemma4-31B. Thus,

judge choice mainly affects ordering inside the middle of the leaderboard rather than the broad high- and low-performing groups.

The ground-truth reference ranks fifth under the strong-judge aggregate, but its rank varies substantially by judge: third under Gemma4-31B, seventh under Qwen3.5-27B, ninth under Qwen3.5-122B-A10, and between eighth and fifteenth for the remaining judges. This variation is consistent with the main interpretation of ELO as comparative plausibility rather than exact reference recovery. Stronger judges often prefer polished large-model continuations to terse human references, while weaker judges do not produce a uniformly higher or lower placement for ground truth.

Figure 6 shows the full pairwise agreement matrix. Under the Qwen3.5-122B-A10 judge, most small models lose clearly to ground truth, while only the strongest large baselines and Gemma4-31B LoRA exceed parity. The Gemma4-31B judge (Figure 7) places ground truth higher (rank 3), so fewer models surpass it. This judge-dependent variation in ground-truth placement further confirms that ELO rankings should not be interpreted as a fixed ground-truth ordering.

Figure 8 shows scale trends for all metrics. Believability and ELO increase most clearly with scale, while TwHIN-BERT cosine similarity is relatively flat above 4B parameters and length closeness is non-monotonic. The strongest scale trend is specific to judged plausibility.

The agreement table compares system-level rankings and pairwise win-rate matrices. We also report a length-bias diagnostic for the three strong judges.

You are an expert evaluator of social-media continuations.

Given a user's action history, the thread they are replying to, and two predicted replies, choose which reply is more believable and logically consistent as a continuation for this user.

Judge only the quality of the continuation for this user and thread. Do not prefer a reply just because it is longer. If both replies are imperfect, choose the better one.

Return ONLY valid JSON with this exact schema:

```
{
  "winner": "A" or "B",
  "brief_rationale": "<one sentence>"
}
```

```
--- ACTION HISTORY (system context) ---
{system_content}
```

```
--- THREAD CONTEXT (user message) ---
{user_content}
```

```
--- REPLY A ---
{reply_a}
```

```
--- REPLY B ---
{reply_b}
```

Figure 4: Pairwise believability prompt used for ELO tournaments. Candidate order is randomized deterministically per example and model pair.

| System                | Agg. rank | Mean rank | SD  | Range | Tier   |
|-----------------------|-----------|-----------|-----|-------|--------|
| Gemma4-31B LoRA       | 1         | 2.5       | 2.1 | 1–8   | upper  |
| Gemma4-31B Base       | 2         | 4.8       | 2.4 | 2–11  | upper  |
| Qwen3.5-122B-A10 Base | 3         | 2.4       | 1.5 | 1–5   | upper  |
| Qwen3.5-27B Base      | 4         | 3.3       | 1.5 | 2–7   | upper  |
| Ground truth          | 5         | 10.2      | 3.5 | 3–15  | middle |
| OLMo3.1-32B Base      | 6         | 6.8       | 3.5 | 4–15  | middle |
| OLMo3.1-32B LoRA      | 7         | 10.8      | 3.9 | 6–18  | middle |
| Qwen3.5-27B LoRA      | 8         | 9.5       | 3.1 | 7–16  | middle |
| Llama3.1-8B LoRA      | 9         | 11.3      | 3.2 | 6–18  | middle |
| Qwen3.5-9B Base       | 10        | 7.0       | 3.5 | 2–15  | middle |
| Gemma4-E4B LoRA       | 11        | 12.5      | 3.1 | 8–18  | middle |
| Qwen3-4B LoRA         | 12        | 12.7      | 3.1 | 9–19  | middle |
| Qwen3.5-9B LoRA       | 13        | 14.0      | 3.6 | 9–21  | middle |
| Qwen3.5-4B LoRA       | 14        | 13.8      | 4.0 | 7–18  | middle |
| Llama3.1-8B Base      | 15        | 17.5      | 3.4 | 12–22 | lower  |
| Qwen3-4B Base         | 16        | 7.5       | 5.0 | 1–17  | middle |
| OLMo3-7B LoRA         | 17        | 17.3      | 4.3 | 6–21  | lower  |
| Gemma4-E4B Base       | 18        | 15.1      | 4.2 | 8–23  | middle |
| Qwen3.5-2B LoRA       | 19        | 18.3      | 3.7 | 12–24 | lower  |
| Qwen3.5-4B Base       | 20        | 20.0      | 2.6 | 14–23 | lower  |
| Gemma4-E2B Base       | 21        | 20.9      | 1.9 | 17–24 | lower  |
| OLMo3-7B Base         | 22        | 17.4      | 4.4 | 9–23  | lower  |
| Qwen3.5-0.8B LoRA     | 23        | 21.8      | 3.1 | 13–24 | lower  |
| Qwen3.5-0.8B Base     | 24        | 25.2      | 0.9 | 24–26 | lower  |
| Gemma4-E2B LoRA       | 25        | 23.6      | 2.6 | 17–26 | lower  |
| Qwen3.5-2B Base       | 26        | 24.8      | 1.4 | 21–26 | lower  |

Table 3: Rank stability across non-LoRA pairwise judges. Mean rank, standard deviation, and range are computed over all judge-specific ELO rankings. Smaller ranks are better. Tiers are descriptive summaries based on mean rank: upper ( $\leq 6$ ), middle ( $> 6$  and  $\leq 16$ ), and lower ( $> 16$ ).

## System rank variation across judges

Darker cells indicate better rank under that judge.

| System                   | Qwen122 | Gemma31 | Qwen27 | OLMo32 | OLMo7 | GemmaE4 8B | GemmaE2 5.2B | Qwen3-4 | Qwen2 | Qwen0.8 |
|--------------------------|---------|---------|--------|--------|-------|------------|--------------|---------|-------|---------|
| 1. Gemma4-31B LoRA       | 1       | 1       | 1      | 1      | 3     | 2          | 4            | 3       | 8     | 1       |
| 2. Gemma4-31B Base       | 4       | 2       | 2      | 5      | 5     | 5          | 5            | 5       | 4     | 11      |
| 3. Qwen3.5-122B-A10 Base | 2       | 5       | 3      | 3      | 2     | 1          | 1            | 1       | 1     | 5       |
| 4. Qwen3.5-27B Base      | 3       | 4       | 4      | 2      | 7     | 3          | 2            | 2       | 2     | 4       |
| 5. Ground truth          | 9       | 3       | 7      | 12     | 11    | 15         | 14           | 14      | 9     | 8       |
| 6. OLMo3.1-32B Base      | 5       | 15      | 5      | 4      | 6     | 4          | 10           | 4       | 5     | 10      |
| 7. OLMo3.1-32B LoRA      | 6       | 11      | 6      | 11     | 10    | 10         | 18           | 7       | 17    | 12      |
| 8. Qwen3.5-27B LoRA      | 8       | 7       | 8      | 7      | 9     | 14         | 16           | 12      | 7     | 7       |
| 9. Llama3.1-8B LoRA      | 12      | 6       | 9      | 14     | 18    | 11         | 8            | 10      | 12    | 13      |
| 10. Qwen3.5-9B Base      | 7       | 8       | 15     | 10     | 4     | 7          | 3            | 8       | 6     | 2       |
| 11. Gemma4-E4B LoRA      | 10      | 14      | 11     | 8      | 15    | 13         | 9            | 11      | 16    | 18      |
| 12. Qwen3-4B LoRA        | 17      | 10      | 10     | 15     | 12    | 19         | 11           | 13      | 11    | 9       |
| 13. Qwen3.5-9B LoRA      | 11      | 16      | 12     | 9      | 16    | 9          | 17           | 21      | 14    | 15      |
| 14. Qwen3.5-4B LoRA      | 16      | 9       | 14     | 16     | 8     | 18         | 7            | 16      | 18    | 16      |
| 15. Llama3.1-8B Base     | 15      | 12      | 16     | 22     | 20    | 16         | 13           | 19      | 22    | 20      |
| 16. Qwen3-4B Base        | 14      | 17      | 13     | 6      | 1     | 6          | 6            | 6       | 3     | 3       |
| 17. OLMo3-7B LoRA        | 18      | 13      | 20     | 20     | 17    | 21         | 19           | 20      | 19    | 6       |
| 18. Gemma4-E4B Base      | 13      | 18      | 18     | 18     | 13    | 8          | 15           | 15      | 10    | 23      |
| 19. Qwen3.5-2B LoRA      | 19      | 19      | 19     | 23     | 24    | 12         | 12           | 18      | 20    | 17      |
| 20. Qwen3.5-4B Base      | 22      | 20      | 17     | 19     | 22    | 23         | 20           | 22      | 21    | 14      |
| 21. Gemma4-E2B Base      | 21      | 22      | 21     | 21     | 19    | 20         | 21           | 17      | 23    | 24      |
| 22. OLMo3-7B Base        | 20      | 23      | 22     | 13     | 14    | 17         | 22           | 9       | 15    | 19      |
| 23. Qwen3.5-0.8B LoRA    | 23      | 21      | 23     | 24     | 23    | 22         | 24           | 23      | 13    | 22      |
| 24. Qwen3.5-0.8B Base    | 26      | 24      | 24     | 26     | 25    | 26         | 25           | 26      | 24    | 26      |
| 25. Gemma4-E2B LoRA      | 25      | 25      | 26     | 17     | 21    | 24         | 23           | 24      | 26    | 25      |
| 26. Qwen3.5-2B Base      | 24      | 26      | 25     | 25     | 26    | 25         | 26           | 25      | 25    | 21      |

Figure 5: System rank variation across non-LoRA pairwise judges. Rows are ordered by the strong-judge aggregate rank; darker cells indicate better rank under the corresponding judge. The figure shows a stable upper tier, a broad middle tier with judge-dependent swaps, and a consistently weak lower tier.

## Judge agreement heatmap

|              | Qwen122 | Gemma31 | Qwen27 | OLMo32 | OLMo7 | GemmaE4 8B | GemmaE2 5.2B | Qwen3-4 | Qwen2 | Qwen0.8 |
|--------------|---------|---------|--------|--------|-------|------------|--------------|---------|-------|---------|
| Qwen122      | 1.00    | 0.66    | 0.80   | 0.72   | 0.67  | 0.76       | 0.61         | 0.72    | 0.64  | 0.57    |
| Gemma31      | 0.66    | 1.00    | 0.74   | 0.51   | 0.53  | 0.48       | 0.62         | 0.53    | 0.51  | 0.56    |
| Qwen27       | 0.80    | 0.74    | 1.00   | 0.68   | 0.61  | 0.61       | 0.59         | 0.64    | 0.59  | 0.56    |
| OLMo32       | 0.72    | 0.51    | 0.68   | 1.00   | 0.70  | 0.69       | 0.56         | 0.73    | 0.62  | 0.56    |
| OLMo7        | 0.67    | 0.53    | 0.61   | 0.70   | 1.00  | 0.61       | 0.62         | 0.68    | 0.66  | 0.58    |
| GemmaE4 8B   | 0.76    | 0.48    | 0.61   | 0.69   | 0.61  | 1.00       | 0.65         | 0.76    | 0.66  | 0.49    |
| GemmaE2 5.2B | 0.61    | 0.62    | 0.59   | 0.56   | 0.62  | 0.65       | 1.00         | 0.66    | 0.59  | 0.51    |
| Qwen3-4      | 0.72    | 0.53    | 0.64   | 0.73   | 0.68  | 0.76       | 0.66         | 1.00    | 0.64  | 0.52    |
| Qwen2        | 0.64    | 0.51    | 0.59   | 0.62   | 0.66  | 0.66       | 0.59         | 0.64    | 1.00  | 0.55    |
| Qwen0.8      | 0.57    | 0.56    | 0.56   | 0.56   | 0.58  | 0.49       | 0.51         | 0.52    | 0.55  | 1.00    |

## Win-matrix agreement

|              | Qwen122 | Gemma31 | Qwen27 | OLMo32 | OLMo7 | GemmaE4 8B | GemmaE2 5.2B | Qwen3-4 | Qwen2 | Qwen0.8 |
|--------------|---------|---------|--------|--------|-------|------------|--------------|---------|-------|---------|
| Qwen122      | 1.00    | 0.93    | 0.95   | 0.91   | 0.87  | 0.92       | 0.85         | 0.90    | 0.86  | 0.83    |
| Gemma31      | 0.93    | 1.00    | 0.95   | 0.87   | 0.84  | 0.87       | 0.83         | 0.86    | 0.85  | 0.82    |
| Qwen27       | 0.95    | 0.95    | 1.00   | 0.89   | 0.85  | 0.89       | 0.85         | 0.88    | 0.85  | 0.82    |
| OLMo32       | 0.91    | 0.87    | 0.89   | 1.00   | 0.93  | 0.94       | 0.89         | 0.94    | 0.90  | 0.87    |
| OLMo7        | 0.87    | 0.84    | 0.85   | 0.93   | 1.00  | 0.91       | 0.92         | 0.92    | 0.92  | 0.89    |
| GemmaE4 8B   | 0.92    | 0.87    | 0.89   | 0.94   | 0.91  | 1.00       | 0.89         | 0.94    | 0.89  | 0.85    |
| GemmaE2 5.2B | 0.85    | 0.83    | 0.85   | 0.89   | 0.92  | 0.89       | 1.00         | 0.89    | 0.91  | 0.89    |
| Qwen3-4      | 0.90    | 0.86    | 0.88   | 0.94   | 0.92  | 0.94       | 0.89         | 1.00    | 0.90  | 0.86    |
| Qwen2        | 0.86    | 0.85    | 0.85   | 0.90   | 0.92  | 0.89       | 0.91         | 0.90    | 1.00  | 0.92    |
| Qwen0.8      | 0.83    | 0.82    | 0.82   | 0.87   | 0.89  | 0.85       | 0.89         | 0.86    | 0.92  | 1.00    |

Figure 6: Agreement among non-LoRA pairwise judges. The left panel compares system-level ELO rankings with Kendall  $\tau$ ; the right panel compares pairwise win-rate matrices. Darker cells indicate higher agreement. Qwen3.5-122B-A10 agrees most strongly with Qwen3.5-27B ( $\tau = 0.803$ ), while agreement with sub-4B judges drops below  $\tau = 0.60$ .

### Model win rate against ground truth (Gemma4-31B judge)

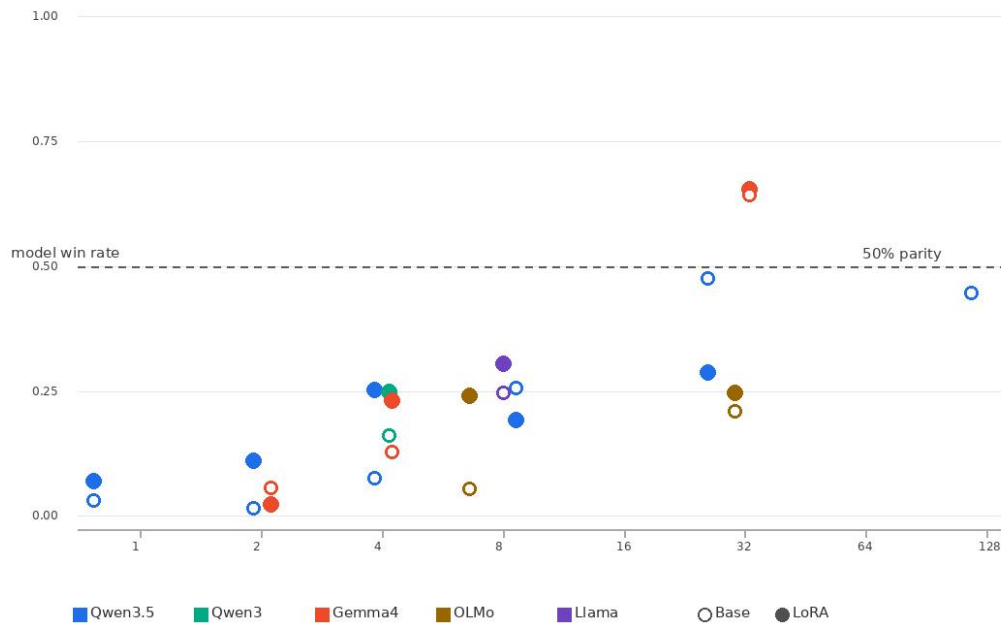


Figure 7: Model win rate against the ground-truth reply under the Gemma4-31B judge. Gemma4-31B places ground truth at rank 3, compared with rank 9 under Qwen3.5-122B-A10, so fewer models exceed parity.

### Model scale trends across metrics

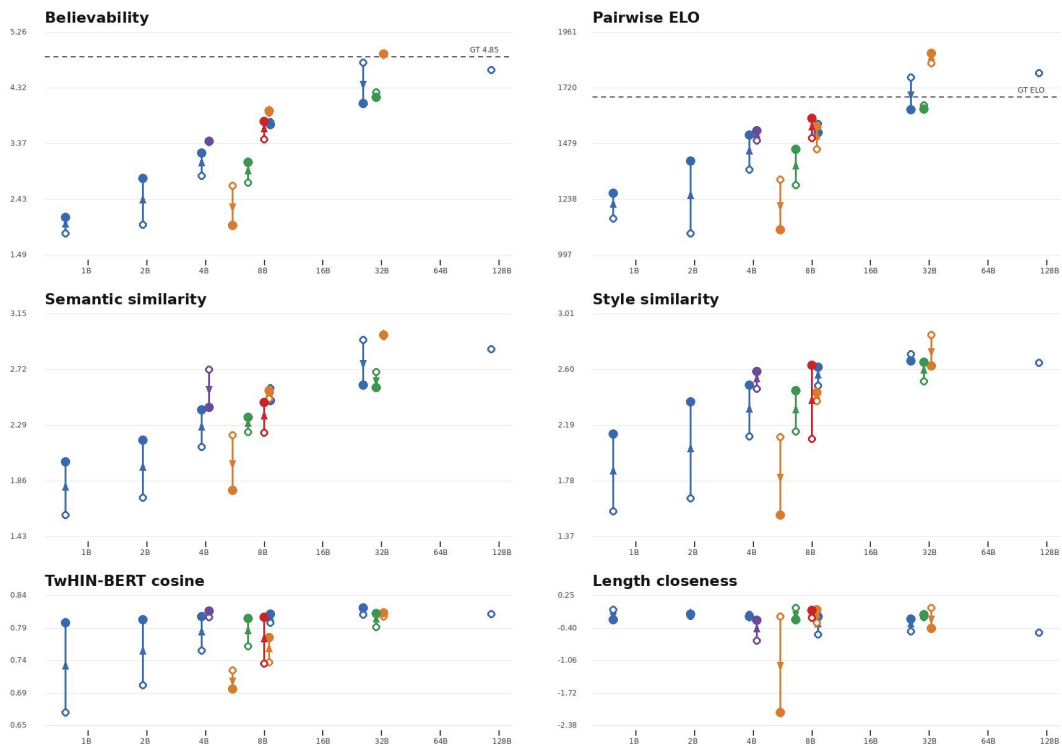


Figure 8: Model scale trends for all main metrics. Believability and ELO are shown in the main text; semantic similarity, style similarity, TwHIN-BERT cosine similarity, and length closeness are included here to show that the scale trend is strongest for judged plausibility and weaker for reference-similarity metrics.

| Judge A          | Judge B      | Kendall $\tau$ | Win-matrix agr. | Pairs |
|------------------|--------------|----------------|-----------------|-------|
| Qwen3.5-122B-A10 | Gemma4-31B   | 0.662          | 0.926           | 325   |
| Qwen3.5-122B-A10 | Qwen3.5-27B  | 0.803          | 0.949           | 325   |
| Qwen3.5-122B-A10 | OLMo3.1-32B  | 0.717          | 0.907           | 325   |
| Qwen3.5-122B-A10 | OLMo3-7B     | 0.668          | 0.868           | 325   |
| Qwen3.5-122B-A10 | Gemma4-E4B   | 0.760          | 0.916           | 325   |
| Qwen3.5-122B-A10 | Gemma4-E2B   | 0.606          | 0.853           | 325   |
| Qwen3.5-122B-A10 | Qwen3-4B     | 0.717          | 0.904           | 325   |
| Qwen3.5-122B-A10 | Qwen3.5-2B   | 0.637          | 0.861           | 325   |
| Qwen3.5-122B-A10 | Qwen3.5-0.8B | 0.569          | 0.827           | 325   |
| Gemma4-31B       | Qwen3.5-27B  | 0.735          | 0.946           | 325   |
| Gemma4-31B       | OLMo3.1-32B  | 0.514          | 0.873           | 325   |
| Gemma4-31B       | OLMo3-7B     | 0.526          | 0.839           | 325   |
| Gemma4-31B       | Gemma4-E4B   | 0.483          | 0.873           | 325   |
| Gemma4-31B       | Gemma4-E2B   | 0.625          | 0.835           | 325   |
| Gemma4-31B       | Qwen3-4B     | 0.526          | 0.865           | 325   |
| Gemma4-31B       | Qwen3.5-2B   | 0.508          | 0.845           | 325   |
| Gemma4-31B       | Qwen3.5-0.8B | 0.563          | 0.816           | 325   |
| Qwen3.5-27B      | OLMo3.1-32B  | 0.680          | 0.890           | 325   |
| Qwen3.5-27B      | OLMo3-7B     | 0.606          | 0.852           | 325   |
| Qwen3.5-27B      | Gemma4-E4B   | 0.612          | 0.891           | 325   |
| Qwen3.5-27B      | Gemma4-E2B   | 0.594          | 0.846           | 325   |
| Qwen3.5-27B      | Qwen3-4B     | 0.643          | 0.883           | 325   |
| Qwen3.5-27B      | Qwen3.5-2B   | 0.588          | 0.854           | 325   |
| Qwen3.5-27B      | Qwen3.5-0.8B | 0.557          | 0.822           | 325   |
| OLMo3.1-32B      | OLMo3-7B     | 0.705          | 0.928           | 325   |
| OLMo3.1-32B      | Gemma4-E4B   | 0.686          | 0.935           | 325   |
| OLMo3.1-32B      | Gemma4-E2B   | 0.557          | 0.892           | 325   |
| OLMo3.1-32B      | Qwen3-4B     | 0.729          | 0.938           | 325   |
| OLMo3.1-32B      | Qwen3.5-2B   | 0.625          | 0.898           | 325   |
| OLMo3.1-32B      | Qwen3.5-0.8B | 0.557          | 0.867           | 325   |
| OLMo3-7B         | Gemma4-E4B   | 0.612          | 0.909           | 325   |
| OLMo3-7B         | Gemma4-E2B   | 0.618          | 0.915           | 325   |
| OLMo3-7B         | Qwen3-4B     | 0.680          | 0.918           | 325   |
| OLMo3-7B         | Qwen3.5-2B   | 0.662          | 0.918           | 325   |
| OLMo3-7B         | Qwen3.5-0.8B | 0.582          | 0.890           | 325   |
| Gemma4-E4B       | Gemma4-E2B   | 0.649          | 0.886           | 325   |
| Gemma4-E4B       | Qwen3-4B     | 0.760          | 0.944           | 325   |
| Gemma4-E4B       | Qwen3.5-2B   | 0.655          | 0.893           | 325   |
| Gemma4-E4B       | Qwen3.5-0.8B | 0.489          | 0.853           | 325   |
| Gemma4-E2B       | Qwen3-4B     | 0.655          | 0.894           | 325   |
| Gemma4-E2B       | Qwen3.5-2B   | 0.588          | 0.912           | 325   |
| Gemma4-E2B       | Qwen3.5-0.8B | 0.508          | 0.890           | 325   |
| Qwen3-4B         | Qwen3.5-2B   | 0.637          | 0.899           | 325   |
| Qwen3-4B         | Qwen3.5-0.8B | 0.520          | 0.864           | 325   |
| Qwen3.5-2B       | Qwen3.5-0.8B | 0.551          | 0.918           | 325   |

Table 4: Agreement between non-LoRA pairwise believability judges. Kendall  $\tau$  compares system-level ELO rankings; win-matrix agreement is one minus the mean absolute difference in pairwise win rates over shared model pairs.

| Judge            | A win | Longer win | $\Delta$ words | Corr.  |
|------------------|-------|------------|----------------|--------|
| Qwen3.5-122B-A10 | 0.575 | 0.575      | -0.1           | 0.001  |
| Gemma4-31B       | 0.524 | 0.553      | -2.0           | -0.049 |
| Qwen3.5-27B      | 0.534 | 0.547      | -2.0           | -0.050 |

Table 5: Verbosity and position-bias diagnostics for the three strong pairwise judges. A win is the fraction of usable comparisons choosing candidate A. Longer win is the fraction of non-tie comparisons where the longer candidate is selected.  $\Delta$  words is winner length minus loser length. Corr. is the correlation between A-minus-B length and choosing A.

| <b>Judge</b>     | <b>Own self win</b> | <b>Strong ref.</b> | <b>Lift</b> | <b>Comparisons</b> |
|------------------|---------------------|--------------------|-------------|--------------------|
| Qwen3.5-122B-A10 | 0.824               | 0.751              | 0.073       | 1100               |
| Gemma4-31B       | 0.880               | 0.855              | 0.026       | 2112               |
| Qwen3.5-27B      | 0.755               | 0.737              | 0.018       | 2106               |
| OLMo3.1-32B      | 0.637               | 0.634              | 0.004       | 2110               |
| OLMo3-7B         | 0.460               | 0.380              | 0.080       | 2111               |
| Gemma4-E4B       | 0.532               | 0.496              | 0.036       | 2112               |
| Gemma4-E2B       | 0.409               | 0.212              | 0.197       | 2109               |
| Qwen3-4B         | 0.634               | 0.511              | 0.122       | 2112               |
| Qwen3.5-2B       | 0.384               | 0.286              | 0.099       | 2107               |
| Qwen3.5-0.8B     | 0.382               | 0.173              | 0.209       | 1300               |

Table 6: Base-plus-LoRA self-preference diagnostic for pairwise judges. The self-label set contains the matching base-model candidate plus its LoRA adapter when available. Own self win is the win rate that judge assigns to its self-label set against all other candidates. Strong ref. is the mean win rate assigned to that same self-label set by the three strong-reference judges used in the main ELO aggregate, excluding the judge itself when applicable. Lift is the difference.

## H Classifier-Based Discriminability

**Setup.** For each generated system, we form a balanced real-versus-generated dataset from reply text. Features combine mean-pooled frozen TwHIN-BERT embeddings with length, punctuation, URL, hashtag, mention, emoji, and formatting statistics. We separate users between training and test and fit logistic-regression and histogram-gradient-boosting classifiers. ROC-AUC is the primary statistic; 0.5 means that real and generated replies are indistinguishable. This test follows prior computational detection work (Pagan et al., 2025; Dawkins et al., 2025). The per-system classifiers measure how readily each generator can be detected; the pooled classifier gives an overall task baseline.

**Controls.** A real-versus-real null gives AUC .51–.52. Length matching preserves the main reductions, including .885 to .629 for Qwen3.5-27B. Repeating user-grouped splits 12 times gives the same family pattern: Qwen, OLMo, and Llama QLoRA AUCs average .65–.73, while all three Gemma

adapters remain more detectable than their bases in at least 11 of 12 splits. A pooled detector over all systems obtains AUC .805 and accuracy .758.

**Context-conditioned test.** The main detector tests surface imitation because it sees reply text only. We therefore add cosine and Jaccard similarity to the user history and stimulus, plus reply length relative to the parent post, and train a detector that sees text and these relational features. In a forced-choice test over a real and generated reply for the same context, the classifier selects the real reply 62–70% of the time for Qwen, OLMo, and Llama QLoRA systems and 93–97% for Gemma. Pooled accuracy is .761. QLoRA makes surface form substantially more human-like for most families, but detectable context-fit differences remain.

**ELO construction and orthogonality.** We obtain out-of-fold probabilities that each reply is human, compare probabilities within each test item, and convert those comparisons to the same ELO scale used for judge rankings. The ladder summarizes classifier outputs and introduces no additional annotations. Across the 24 generated systems, detector AUC has no detectable rank association with scalar believability ( $\rho = .02$ ,  $p = .93$ , bootstrap 95% CI  $[-.44, .48]$ ) or judge ELO ( $\rho = -.05$ ,  $p = .83$ ). The judge and detector ladder ranks correlate at  $\rho = .28$  when human replies are included ( $p = .18$ ,  $n = 25$ ).

| Model            | $ \theta (\mathbf{B})$ | base<br>AUC | LoRA<br>AUC | base <sub><math>\ell</math></sub><br>AUC | LoRA <sub><math>\ell</math></sub><br>AUC | $\Delta$<br>(base—LoRA) |
|------------------|------------------------|-------------|-------------|------------------------------------------|------------------------------------------|-------------------------|
| Qwen3.5-0.8B     | 0.8                    | 0.982       | 0.726       | 0.969                                    | 0.693                                    | 0.256                   |
| Qwen3.5-2B       | 2                      | 0.972       | 0.644       | 0.937                                    | 0.674                                    | 0.328                   |
| Qwen3-4B         | 4                      | —           | 0.663       | —                                        | 0.647                                    | —                       |
| Qwen3.5-4B       | 4                      | 0.935       | 0.659       | 0.868                                    | 0.659                                    | 0.276                   |
| Gemma4-E2B       | 5.2                    | 0.895       | 0.954       | 0.863                                    | 0.883                                    | -0.059                  |
| OLMo3-7B         | 7                      | 0.981       | 0.717       | 0.980                                    | 0.671                                    | 0.264                   |
| Gemma4-E4B       | 8                      | 0.925       | 0.939       | 0.909                                    | 0.885                                    | -0.014                  |
| Llama3.1-8B      | 8                      | 0.973       | 0.703       | 0.964                                    | 0.675                                    | 0.270                   |
| Qwen3.5-9B       | 9                      | 0.911       | 0.654       | 0.857                                    | 0.671                                    | 0.257                   |
| Qwen3.5-27B      | 27                     | 0.921       | 0.643       | 0.885                                    | 0.629                                    | 0.279                   |
| Gemma4-31B       | 31                     | 0.926       | 0.983       | 0.885                                    | 0.988                                    | -0.057                  |
| OLMo3.1-32B      | 32                     | 0.961       | 0.672       | 0.970                                    | 0.708                                    | 0.289                   |
| Qwen3.5-122B-A10 | 122                    | 0.982       | —           | 0.970                                    | —                                        | —                       |

Table 7: Per-system detector ROC-AUC. The subscript  $\ell$  denotes reply-length-matched data. Lower values indicate that generated replies are more difficult to separate from real replies.  $\Delta$  is raw base AUC minus QLoRA AUC.

| System                      | judge ELO | rank | detector ELO | rank |
|-----------------------------|-----------|------|--------------|------|
| <i>human (ground truth)</i> | 1680      | 5    | 1673         | 1    |
| Qwen3.5-27B (L)             | 1626      | 8    | 1621         | 2    |
| OLMo3.1-32B (L)             | 1628      | 7    | 1619         | 3    |
| Qwen3.5-9B (L)              | 1527      | 13   | 1616         | 4    |
| Qwen3-4B (L)                | 1534      | 12   | 1615         | 5    |
| Llama3.1-8B (L)             | 1588      | 9    | 1612         | 6    |
| OLMo3-7B (L)                | 1454      | 17   | 1609         | 7    |
| Qwen3.5-4B (L)              | 1515      | 14   | 1600         | 8    |
| Qwen3.5-2B (L)              | 1403      | 19   | 1598         | 9    |
| Qwen3.5-0.8B (L)            | 1265      | 23   | 1569         | 10   |
| OLMo3.1-32B (B)             | 1644      | 6    | 1489         | 11   |
| Qwen3.5-9B (B)              | 1562      | 10   | 1487         | 12   |
| Qwen3.5-27B (B)             | 1764      | 4    | 1486         | 13   |
| Gemma4-31B (B)              | 1825      | 2    | 1477         | 14   |
| Llama3.1-8B (B)             | 1503      | 15   | 1477         | 15   |
| Qwen3.5-4B (B)              | 1366      | 20   | 1464         | 16   |
| Gemma4-E2B (B)              | 1323      | 21   | 1460         | 17   |
| Gemma4-E4B (B)              | 1452      | 18   | 1439         | 18   |
| OLMo3-7B (B)                | 1298      | 22   | 1431         | 19   |
| Qwen3.5-2B (B)              | 1090      | 26   | 1424         | 20   |
| Gemma4-E4B (L)              | 1552      | 11   | 1423         | 21   |
| Gemma4-31B (L)              | 1867      | 1    | 1389         | 22   |
| Qwen3.5-0.8B (B)            | 1153      | 24   | 1381         | 23   |
| Gemma4-E2B (L)              | 1105      | 25   | 1304         | 24   |
| Qwen3.5-122B-A10 (B)        | 1783      | 3    | 1235         | 25   |

Table 8: Contextual-believability ELO from the strong LLM judges and discriminability ELO from cross-fitted classifier scores. B and L denote base and QLoRA systems.

## I Expanded Human Audit

**Items and raters.** Two paper authors independently rated the same 200 comparisons, giving 400 ratings. Six external, non-author raters each completed the same 30-comparison validation set, giving 180 ratings. The external raters were undergraduate and graduate students recruited from the authors’ research lab; none was involved in this project. Twenty-nine external-audit comparisons also appear in the author set. We report the two rater pools separately unless an analysis is explicitly described as pooled.

**Sampling strata.** For sampling only, we used a broader six-judge pool comprising Qwen3.5-122B-A10, Gemma4-31B, Qwen3.5-27B, OLMo3.1-32B, GPT-5.4, and GPT-5.4-mini to identify unanimous and contested strata. This pool does not define a paper ranking. All human-judge agreement results below use the three judges in the main ELO aggregate: Qwen3.5-122B-A10, Gemma4-31B, and Qwen3.5-27B.

The author set contains 100 real-generated cases where Qwen3.5-122B-A10 selected the generated reply, 42 contested model-model pairs, 12 model-model pairs on which the sampling pool was unanimous, 10 clear model pairs with a large ELO gap, 30 real-generated pairs where the pool selected the real reply, and 6 real-generated pairs where it split. The adversarial cases contain 71 temporal-holdout and 29 held-out-user examples. The external set

| Author-audit stratum       | Items | Ties  | Agree | $\kappa$ |
|----------------------------|-------|-------|-------|----------|
| Adversarial real-generated | 100   | 29.0% | .42   | .12      |
| Clear model pairs          | 10    | 0.0%  | .90   | .80      |
| Unanimous model pairs      | 12    | 8.3%  | .58   | .28      |
| Contested model pairs      | 42    | 21.4% | .55   | .30      |
| Pool selected real         | 30    | 5.0%  | .60   | .27      |
| Pool split on real         | 6     | 0.0%  | .67   | .33      |

Table 9: Inter-annotator agreement for the 200 comparisons rated by both authors. “Pool” denotes the six-judge pool used only to construct sampling strata. Ties are percentages of individual votes; agreement and Cohen’s  $\kappa$  treat ratings as A, B, or tie.

contains 10 contested model pairs, 4 model pairs selected unanimously by the sampling pool, 3 clear model pairs, and 13 real-generated pairs: 4 where the pool selected the real reply, 4 where it split, and 5 adversarial generated wins. These strata compare difficulty levels; they do not estimate their frequency in the full tournament.

**Blinding and question.** Every rater saw four fields in order: “User History” (past posts and replies), “Thread / Stimulus,” “Response A,” and “Response B.” Candidate order was randomized; system identities, judge decisions, and the real reply were hidden.

**Instructions shown to raters.** The authors and raters were asked, “Which response is the more believable reply from THIS specific user in THIS thread?” The choices were “A much better,” “A slightly better,” “Tie / unclear,” “B slightly better,” and “B much better.” They were instructed to judge “voice, tone, length habits, formatting quirks, stance, and fit to the thread” and to ask “would this user have written this?”, not “which reply is nicer or better written?” They were told to use ties only when they genuinely could not decide, to work alone, and not to share the displayed text.

**Author agreement follows difficulty.** Across the five stratified categories (100 comparisons), exact three-way agreement is .61 and Cohen’s  $\kappa = .35$ . After adding the adversarial cases, agreement across all 200 comparisons is .515 and  $\kappa = .24$ . On the 141 comparisons where both authors choose a side, directional agreement is .574. These aggregates hide the intended difficulty gradient in Table 9: agreement is .90 ( $\kappa = .80$ ) on clear pairs, .55–.67 ( $\kappa = .27$ –.33) on the four intermediate strata, and .42 ( $\kappa = .12$ ) on adversarial real-generated cases. The tie rate likewise rises from 11.5% across the five stratified categories to

| Adversarial ratings                        | Generated | Real  | Tie   |
|--------------------------------------------|-----------|-------|-------|
| Authors, all cases ( $n = 200$ votes)      | 41.5%     | 29.5% | 29.0% |
| External, shared cases ( $n = 30$ votes)   | 46.7%     | 43.3% | 10.0% |
| All raters, shared cases ( $n = 40$ votes) | 42.5%     | 42.5% | 15.0% |

Table 10: Choices on cases selected because Qwen3.5-122B-A10 preferred a generated reply to the real reply. The first row covers all 100 such author-audit comparisons. The last two rows cover the five cases rated by all eight raters. Percentages include ties.

29.0% on the adversarial cases.

Cohen’s  $\kappa$  treats a tie from one author and a directional choice from the other as disagreement. If such cases are treated as abstentions and removed,  $\kappa$  rises from .12 to .32 on the adversarial slice and from .35 to .44 across the five stratified categories. We retain all comparisons in the primary results. Under either treatment, low agreement is concentrated in comparisons selected to be difficult.

#### Ground-truth losses are ambiguous to humans.

Table 10 reports all votes rather than dropping ties. Although every case was selected because the judge preferred generated text, only 41.5% of author ratings made the same choice; 29.5% preferred the real reply and 29.0% could not distinguish them. On the original five-point scale, 16.5% rated the generated reply much better and 25.0% slightly better; 20.5% rated the real reply slightly better and only 9.0% much better. Thus, only 9.0% of votes strongly reverse the judge. Among decisive author votes, generated replies receive 58.5%, but that conditional percentage omits nearly one-third of the evidence.

The external ratings show the same uncertainty. On their five adversarial cases, generated and real replies receive 46.7% and 43.3% of votes, respectively, and the external majority selects the real reply on only one of five comparisons. Pooling all eight raters on these five cases produces an exact 42.5%–42.5% split, with 15.0% ties. Thus, ground-truth losses are neither straightforward judge failures nor clear human endorsements of generated text.

**External raters.** Across all 13 real–generated comparisons in the external audit, raters select the real reply in 42/78 votes (53.8%), the generated reply in 30/78 (38.5%), and tie in 6/78 (7.7%). This is only 58.3% real among decisive votes, showing that identifying the observed reply remains difficult even when items are not restricted to adversarial judge decisions. Across the full 30 comparisons,

the external tie rate is 10.0%; mean pairwise three-way agreement is .50, mean pairwise  $\kappa = .15$ , and Fleiss’  $\kappa = .14$ . Nevertheless, 22/30 comparisons have a majority margin of at least two votes. The low pairwise agreement reflects variation in individual choices and does not imply that every item lacks a stable majority.

The external audit also checks whether using authors changed the result. On the shared comparisons, each author’s leave-one-out agreement with the other raters’ majority is .62 and .67. The corresponding range among external raters is .52–.64; mean author–external  $\kappa$  is .19, compared with .15 among external pairs. We therefore find no evidence that the author ratings systematically depart from the non-author ratings. The shared sample is small, so this is a validation of the reported patterns rather than a population estimate.

**Human and judge agreement.** For the three judges used in the paper’s main ELO aggregate, individual decisions across both rater pools match the judge majority in 275/481 decisive ratings (57.2%), close to the authors’ 57.4% directional agreement when both choose a side. Agreement is 173/263 (65.8%) when all three judges select the same candidate and 102/218 (46.8%) when they split. Among comparisons with a decisive pooled human majority, that majority matches the judges on 55/68 unanimous cases (80.9%) and 30/55 split cases (54.5%). Human preference also responds to the size of the judge–ELO difference. Across 198 decisive model–model ratings from the two authors and six external raters, humans select the higher-ELO system in 42/94 ratings (45%) when the gap is below 150 points, 44/73 (60%) at 150–300 points, and 21/31 (68%) above 300 points. Judge unanimity and ELO distance therefore provide useful confidence signals, but neither determines the human answer.

**Illustrative audit cases.** Table 11 shows two adversarial cases where a strong judge preferred a machine reply to the user’s observed reply. In both, the judge rewards fluency and completeness while penalizing terse or fragmentary human writing.

## J Qualitative Analysis of the 27B Reversal

We inspected the 44 Qwen3.5-27B base-versus-QLoRA comparisons and their GPT-5.4 judge rationales. GPT-5.4 was used through the OpenAI API (OpenAI, 2026). The base wins 27 comparisons

| Thread                          | Observed user reply                                                                                     | Preferred machine reply                                                                                  | Condensed judge rationale                                                                                                                                              |
|---------------------------------|---------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Canadian and US healthcare      | Mentions Saskatchewan health insurance as inspiration for Canada’s system, but the sentence trails off. | Credits Tommy Douglas and advocates for universal coverage while contrasting Canadian and US healthcare. | The machine reply is grammatical and extends the comparison, while the observed reply is incomplete.                                                                   |
| Ownership of the Canadian press | States briefly that American entities own most of the Canadian press.                                   | Argues that mainstream coverage leans centre-right and emphasizes caution and verification.              | The machine reply gives a nuanced opinion consistent with the user’s stated habits, while the observed reply makes a questionable ownership claim without explanation. |

Table 11: Two adversarial audit cases where a strong judge preferred a machine reply over the observed user reply. The judge favored fluent, complete text over terse or fragmentary human writing. Reply texts are paraphrased and rationales condensed to reduce deanonymization risk.

and QLoRA wins 17. The rationales are post-hoc explanations, so the counts below are descriptive and do not establish causes.

When QLoRA loses, rationales most often cite unsupported specifics or implausible claims (7 cases), fragments or awkward coherence (5), off-thread drift (5), and overly generic or polished text (4). When the base loses, generic, polished, or over-long advocacy is the most common negative theme (5 cases). Style-match language praising the winner appears on both sides. The pattern clarifies the 27B metric reversal: QLoRA often reproduces the user’s terse register, but it also adopts risky habits such as unhedged details and loose thread logic. This makes the outputs harder for the text classifier to distinguish from real replies while reducing contextual-believability scores.

## K Full Fine-Tuning vs. QLoRA

We compare epoch-10 QLoRA with epoch-10 full fine-tuning for two smaller models to test whether updating all parameters improves the user-simulation objective beyond parameter-efficient adaptation.

For OLMo3-7B, full fine-tuning and QLoRA produce nearly identical results: believability is 3.123 vs. 3.053, semantic similarity is tied at 2.35, and length ratios are within 0.01 of each other. The two methods converge to a similar solution, suggesting that the LoRA rank-32 bottleneck does not limit adaptation for this model.

For Qwen3-4B, full fine-tuning is also close to QLoRA: believability decreases by 0.008 (3.398 vs. 3.406), style is 2.611 vs. 2.582, and the length ratio is 1.251 vs. 1.253.

For these models, updating all parameters provides little benefit over QLoRA at the matched

checkpoint.

| Thread                     | Base reply                                                                                    | QLoRA reply                                                                                                              | GPT-5.4 decision                                                                                    |
|----------------------------|-----------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| Security clearance         | Gives coherent but generic advocacy about trust and foreign influence, ending with hash-tags. | Copies the user’s terse questions and unusual spacing before punctuation, but claims that refusing clearance is illegal. | Base wins because the QLoRA reply introduces an implausible legal claim.                            |
| NDP leadership             | Gives a long, hedged discussion of centrist versus ideological leadership.                    | Names a specific candidate and confidently supplies unsupported biographical details.                                    | Base wins because the specific candidate and details are less grounded in the user’s history.       |
| Voting and economic policy | Gives a long anti-neoliberal argument using several recycled phrases.                         | States the same position in two terse paragraphs that better match the user’s cadence.                                   | QLoRA wins because the base reply is overly long and less consistent with the user’s concise style. |

Table 12: Representative Qwen3.5-27B base–QLoRA comparisons. Replies and rationales are paraphrased to reduce deanonymization risk.

| Model    | Condition        | Bel.  | Sem.  | Style | Cos.  | Len.  |
|----------|------------------|-------|-------|-------|-------|-------|
| Qwen3-4B | Base             | 3.416 | 2.721 | 2.456 | 0.803 | 1.672 |
| Qwen3-4B | QLoRA epoch 10   | 3.406 | 2.428 | 2.582 | 0.812 | 1.253 |
| Qwen3-4B | Full FT epoch 10 | 3.398 | 2.359 | 2.611 | 0.810 | 1.251 |
| OLMo3-7B | Base             | 2.710 | 2.239 | 2.141 | 0.762 | 1.002 |
| OLMo3-7B | QLoRA epoch 10   | 3.053 | 2.352 | 2.440 | 0.802 | 1.246 |
| OLMo3-7B | Full FT epoch 10 | 3.123 | 2.351 | 2.402 | 0.801 | 1.255 |

Table 13: Epoch-10 full fine-tuning compared with the matched base model and epoch-10 QLoRA adapter at evaluation context 2048. Sem. is the average of content and stance similarity. Len. is the predicted-to-gold length ratio (1.0 is ideal).

## **L Main Metric Uncertainty and Splits**

The scalar intervals and cohort splits qualify the aggregate comparisons. The direction of the QLoRA believability effect agrees between future replies from seen users and replies from held-out users for 11 of 12 matched systems. Gemma4-31B is the sole exception, with near-zero changes in opposite directions ( $-0.023$  and  $+0.032$ ). The capacity-dependent result is therefore not driven by either evaluation cohort alone.

| Model            | Variant | Bel. 95% CI          | Sem. 95% CI          | Style 95% CI         | ELO range         |
|------------------|---------|----------------------|----------------------|----------------------|-------------------|
| Qwen3.5-0.8B     | Base    | 1.854 [1.778, 1.930] | 1.601 [1.551, 1.651] | 1.556 [1.504, 1.607] | 1153 [1084, 1206] |
| Qwen3.5-0.8B     | LoRA    | 2.122 [2.048, 2.196] | 2.008 [1.948, 2.068] | 2.123 [2.058, 2.188] | 1265 [1232, 1314] |
| Qwen3.5-2B       | Base    | 1.992 [1.917, 2.067] | 1.732 [1.680, 1.784] | 1.650 [1.598, 1.703] | 1090 [1058, 1130] |
| Qwen3.5-2B       | LoRA    | 2.780 [2.696, 2.863] | 2.175 [2.114, 2.236] | 2.357 [2.294, 2.421] | 1403 [1375, 1437] |
| Qwen3.5-4B       | Base    | 2.820 [2.733, 2.908] | 2.123 [2.062, 2.184] | 2.103 [2.040, 2.166] | 1366 [1325, 1446] |
| Qwen3.5-4B       | LoRA    | 3.213 [3.127, 3.300] | 2.411 [2.347, 2.475] | 2.481 [2.414, 2.548] | 1515 [1459, 1571] |
| Qwen3.5-9B       | Base    | 3.730 [3.645, 3.815] | 2.578 [2.515, 2.640] | 2.475 [2.411, 2.539] | 1562 [1487, 1624] |
| Qwen3.5-9B       | LoRA    | 3.698 [3.614, 3.782] | 2.484 [2.419, 2.548] | 2.613 [2.548, 2.678] | 1527 [1511, 1538] |
| Qwen3.5-27B      | Base    | 4.737 [4.691, 4.784] | 2.949 [2.888, 3.011] | 2.708 [2.642, 2.775] | 1764 [1727, 1821] |
| Qwen3.5-27B      | LoRA    | 4.046 [3.970, 4.123] | 2.602 [2.537, 2.667] | 2.657 [2.590, 2.723] | 1626 [1603, 1656] |
| Qwen3.5-122B-A10 | Base    | 4.619 [4.566, 4.672] | 2.877 [2.816, 2.938] | 2.643 [2.580, 2.707] | 1783 [1723, 1848] |
| Qwen3-4B         | Base    | 3.416 [3.324, 3.508] | 2.721 [2.657, 2.785] | 2.456 [2.390, 2.522] | 1492 [1473, 1514] |
| Qwen3-4B         | LoRA    | 3.406 [3.321, 3.492] | 2.428 [2.365, 2.491] | 2.582 [2.517, 2.648] | 1534 [1452, 1584] |
| Gemma4-E2B       | Base    | 2.652 [2.555, 2.750] | 2.214 [2.149, 2.278] | 2.101 [2.035, 2.168] | 1323 [1273, 1362] |
| Gemma4-E2B       | LoRA    | 1.990 [1.907, 2.073] | 1.789 [1.731, 1.847] | 1.528 [1.472, 1.585] | 1105 [1082, 1119] |
| Gemma4-E4B       | Base    | 3.889 [3.803, 3.974] | 2.498 [2.435, 2.561] | 2.365 [2.298, 2.432] | 1452 [1427, 1498] |
| Gemma4-E4B       | LoRA    | 3.922 [3.839, 4.006] | 2.559 [2.497, 2.621] | 2.427 [2.363, 2.492] | 1552 [1544, 1560] |
| Gemma4-31B       | Base    | 4.894 [4.864, 4.925] | 2.979 [2.917, 3.041] | 2.848 [2.784, 2.912] | 1825 [1794, 1861] |
| Gemma4-31B       | LoRA    | 4.890 [4.860, 4.921] | 2.987 [2.925, 3.048] | 2.623 [2.561, 2.685] | 1867 [1834, 1898] |
| OLMo3-7B         | Base    | 2.710 [2.629, 2.791] | 2.239 [2.187, 2.290] | 2.141 [2.089, 2.193] | 1298 [1268, 1336] |
| OLMo3-7B         | LoRA    | 3.053 [2.965, 3.140] | 2.352 [2.289, 2.415] | 2.440 [2.374, 2.505] | 1454 [1363, 1561] |
| OLMo3.1-32B      | Base    | 4.236 [4.166, 4.306] | 2.700 [2.644, 2.756] | 2.509 [2.451, 2.568] | 1644 [1530, 1717] |
| OLMo3.1-32B      | LoRA    | 4.159 [4.084, 4.234] | 2.582 [2.514, 2.649] | 2.648 [2.581, 2.715] | 1628 [1566, 1684] |
| Llama3.1-8B      | Base    | 3.447 [3.359, 3.535] | 2.234 [2.176, 2.293] | 2.087 [2.029, 2.145] | 1503 [1467, 1566] |
| Llama3.1-8B      | LoRA    | 3.744 [3.663, 3.825] | 2.467 [2.401, 2.532] | 2.626 [2.559, 2.693] | 1588 [1514, 1635] |

Table 14: 95% confidence intervals for main scalar metrics and strong-judge ELO ranges. Scalar intervals use normal mean intervals over test examples; ELO ranges are the minimum and maximum ELO assigned by the three strong pairwise judges.

| Model            | Variant | Content | Stance | Style | Bel.  | Cos.  | Len.  | Lev.  | Jacc. |
|------------------|---------|---------|--------|-------|-------|-------|-------|-------|-------|
| Qwen3.5-0.8B     | Base    | 1.425   | 1.777  | 1.556 | 1.854 | 0.665 | 0.956 | 0.102 | 0.065 |
| Qwen3.5-0.8B     | LoRA    | 1.781   | 2.236  | 2.123 | 2.122 | 0.795 | 1.246 | 0.164 | 0.127 |
| Qwen3.5-2B       | Base    | 1.495   | 1.969  | 1.650 | 1.992 | 0.705 | 0.835 | 0.114 | 0.074 |
| Qwen3.5-2B       | LoRA    | 1.906   | 2.444  | 2.357 | 2.780 | 0.800 | 1.134 | 0.170 | 0.131 |
| Qwen3.5-4B       | Base    | 1.810   | 2.436  | 2.103 | 2.820 | 0.755 | 1.146 | 0.143 | 0.107 |
| Qwen3.5-4B       | LoRA    | 2.101   | 2.721  | 2.481 | 3.213 | 0.804 | 1.174 | 0.181 | 0.147 |
| Qwen3.5-9B       | Base    | 2.167   | 2.988  | 2.475 | 3.730 | 0.795 | 1.542 | 0.151 | 0.126 |
| Qwen3.5-9B       | LoRA    | 2.123   | 2.844  | 2.613 | 3.698 | 0.808 | 1.174 | 0.176 | 0.144 |
| Qwen3.5-27B      | Base    | 2.475   | 3.424  | 2.708 | 4.737 | 0.806 | 1.476 | 0.165 | 0.145 |
| Qwen3.5-27B      | LoRA    | 2.232   | 2.972  | 2.657 | 4.046 | 0.817 | 1.233 | 0.185 | 0.153 |
| Qwen3.5-122B-A10 | Base    | 2.419   | 3.334  | 2.643 | 4.619 | 0.808 | 1.509 | 0.153 | 0.133 |
| Qwen3-4B         | Base    | 2.289   | 3.153  | 2.456 | 3.416 | 0.803 | 1.672 | 0.157 | 0.139 |
| Qwen3-4B         | LoRA    | 2.118   | 2.737  | 2.582 | 3.406 | 0.812 | 1.253 | 0.174 | 0.142 |
| Gemma4-E2B       | Base    | 1.868   | 2.559  | 2.101 | 2.652 | 0.726 | 0.818 | 0.146 | 0.108 |
| Gemma4-E2B       | LoRA    | 1.544   | 2.034  | 1.528 | 1.990 | 0.699 | 3.122 | 0.110 | 0.098 |
| Gemma4-E4B       | Base    | 2.051   | 2.945  | 2.365 | 3.889 | 0.737 | 0.699 | 0.159 | 0.114 |
| Gemma4-E4B       | LoRA    | 2.111   | 3.006  | 2.427 | 3.922 | 0.774 | 1.046 | 0.158 | 0.117 |
| Gemma4-31B       | Base    | 2.444   | 3.513  | 2.848 | 4.894 | 0.804 | 0.994 | 0.179 | 0.141 |
| Gemma4-31B       | LoRA    | 2.478   | 3.495  | 2.623 | 4.890 | 0.809 | 1.420 | 0.159 | 0.134 |
| OLMo3-7B         | Base    | 1.816   | 2.661  | 2.141 | 2.710 | 0.762 | 1.002 | 0.109 | 0.069 |
| OLMo3-7B         | LoRA    | 2.034   | 2.671  | 2.440 | 3.053 | 0.802 | 1.246 | 0.169 | 0.136 |
| OLMo3.1-32B      | Base    | 2.184   | 3.216  | 2.509 | 4.236 | 0.789 | 1.196 | 0.122 | 0.090 |
| OLMo3.1-32B      | LoRA    | 2.246   | 2.917  | 2.648 | 4.159 | 0.809 | 1.143 | 0.186 | 0.151 |
| Llama3.1-8B      | Base    | 1.850   | 2.618  | 2.087 | 3.447 | 0.736 | 0.801 | 0.130 | 0.088 |
| Llama3.1-8B      | LoRA    | 2.132   | 2.802  | 2.626 | 3.744 | 0.803 | 1.055 | 0.182 | 0.144 |

Table 15: Full scalar and non-LLM metric table for the main 2,048-token comparison.

| Model            | Variant | Cohort       | n   | Bel.  | Sem.  | Style |
|------------------|---------|--------------|-----|-------|-------|-------|
| Qwen3.5-0.8B     | Base    | temporal     | 834 | 1.770 | 1.671 | 1.608 |
| Qwen3.5-0.8B     | Base    | holdout_test | 433 | 2.016 | 1.476 | 1.462 |
| Qwen3.5-0.8B     | LoRA    | temporal     | 838 | 2.158 | 2.097 | 2.180 |
| Qwen3.5-0.8B     | LoRA    | holdout_test | 438 | 2.055 | 1.839 | 2.014 |
| Qwen3.5-2B       | Base    | temporal     | 838 | 1.967 | 1.789 | 1.700 |
| Qwen3.5-2B       | Base    | holdout_test | 437 | 2.041 | 1.624 | 1.556 |
| Qwen3.5-2B       | LoRA    | temporal     | 837 | 2.853 | 2.278 | 2.431 |
| Qwen3.5-2B       | LoRA    | holdout_test | 438 | 2.639 | 1.977 | 2.212 |
| Qwen3.5-4B       | Base    | temporal     | 837 | 2.851 | 2.239 | 2.204 |
| Qwen3.5-4B       | Base    | holdout_test | 438 | 2.763 | 1.904 | 1.911 |
| Qwen3.5-4B       | LoRA    | temporal     | 838 | 3.309 | 2.531 | 2.593 |
| Qwen3.5-4B       | LoRA    | holdout_test | 438 | 3.030 | 2.182 | 2.267 |
| Qwen3.5-9B       | Base    | temporal     | 837 | 3.828 | 2.676 | 2.538 |
| Qwen3.5-9B       | Base    | holdout_test | 438 | 3.543 | 2.390 | 2.356 |
| Qwen3.5-9B       | LoRA    | temporal     | 838 | 3.780 | 2.563 | 2.697 |
| Qwen3.5-9B       | LoRA    | holdout_test | 438 | 3.541 | 2.332 | 2.452 |
| Qwen3.5-27B      | Base    | temporal     | 838 | 4.748 | 3.053 | 2.786 |
| Qwen3.5-27B      | Base    | holdout_test | 438 | 4.717 | 2.752 | 2.559 |
| Qwen3.5-27B      | LoRA    | temporal     | 838 | 4.126 | 2.684 | 2.757 |
| Qwen3.5-27B      | LoRA    | holdout_test | 438 | 3.893 | 2.444 | 2.466 |
| Qwen3.5-122B-A10 | Base    | temporal     | 838 | 4.612 | 2.936 | 2.702 |
| Qwen3.5-122B-A10 | Base    | holdout_test | 438 | 4.632 | 2.764 | 2.532 |
| Qwen3-4B         | Base    | temporal     | 838 | 3.456 | 2.837 | 2.568 |
| Qwen3-4B         | Base    | holdout_test | 438 | 3.340 | 2.500 | 2.242 |
| Qwen3-4B         | LoRA    | temporal     | 838 | 3.446 | 2.511 | 2.647 |
| Qwen3-4B         | LoRA    | holdout_test | 437 | 3.330 | 2.270 | 2.460 |
| Gemma4-E2B       | Base    | temporal     | 833 | 2.562 | 2.274 | 2.173 |
| Gemma4-E2B       | Base    | holdout_test | 436 | 2.826 | 2.102 | 1.963 |
| Gemma4-E2B       | LoRA    | temporal     | 836 | 1.914 | 1.807 | 1.504 |
| Gemma4-E2B       | LoRA    | holdout_test | 437 | 2.135 | 1.745 | 1.561 |
| Gemma4-E4B       | Base    | temporal     | 837 | 3.964 | 2.585 | 2.477 |
| Gemma4-E4B       | Base    | holdout_test | 437 | 3.744 | 2.339 | 2.158 |
| Gemma4-E4B       | LoRA    | temporal     | 838 | 3.970 | 2.634 | 2.498 |
| Gemma4-E4B       | LoRA    | holdout_test | 438 | 3.831 | 2.416 | 2.292 |
| Gemma4-31B       | Base    | temporal     | 838 | 4.894 | 3.048 | 2.905 |
| Gemma4-31B       | Base    | holdout_test | 438 | 4.895 | 2.846 | 2.740 |
| Gemma4-31B       | LoRA    | temporal     | 838 | 4.871 | 3.061 | 2.669 |
| Gemma4-31B       | LoRA    | holdout_test | 437 | 4.927 | 2.849 | 2.535 |
| OLMo3-7B         | Base    | temporal     | 837 | 2.760 | 2.302 | 2.190 |
| OLMo3-7B         | Base    | holdout_test | 437 | 2.616 | 2.116 | 2.046 |
| OLMo3-7B         | LoRA    | temporal     | 837 | 3.111 | 2.432 | 2.532 |
| OLMo3-7B         | LoRA    | holdout_test | 436 | 2.940 | 2.200 | 2.264 |
| OLMo3.1-32B      | Base    | temporal     | 837 | 4.275 | 2.762 | 2.559 |
| OLMo3.1-32B      | Base    | holdout_test | 438 | 4.162 | 2.586 | 2.418 |
| OLMo3.1-32B      | LoRA    | temporal     | 838 | 4.165 | 2.675 | 2.728 |
| OLMo3.1-32B      | LoRA    | holdout_test | 438 | 4.148 | 2.402 | 2.495 |
| Llama3.1-8B      | Base    | temporal     | 837 | 3.435 | 2.314 | 2.157 |
| Llama3.1-8B      | Base    | holdout_test | 438 | 3.470 | 2.086 | 1.957 |
| Llama3.1-8B      | LoRA    | temporal     | 837 | 3.763 | 2.590 | 2.735 |
| Llama3.1-8B      | LoRA    | holdout_test | 438 | 3.708 | 2.228 | 2.420 |

Table 16: Split breakdown for the main scalar metrics. Temporal examples are future replies from users seen in training; holdout examples are replies from held-out users.

## **M Evaluation Context Length**

For Qwen3.5-27B, Gemma4-31B, and OLMo3.1-32B, increasing the evaluation history from 512 to 2048 tokens improves believability for both baselines and LoRA adapters. Gains from 2048 to 6144 are smaller and model-specific. Gemma4-31B continues to improve across most scalar metrics, Qwen3.5-27B mostly saturates on semantic similarity while still gaining believability and style, and OLMo3.1-32B gains believability but shows some length-ratio drift for the base model. Additional user history is most useful when moving from short to moderate context; returns diminish once the prompt contains enough behavioral evidence.

| Model       | Variant | Eval ctx | Bel.  | Sem.  | Style | Cos.  | Len.  |
|-------------|---------|----------|-------|-------|-------|-------|-------|
| Qwen3.5-27B | Base    | 512      | 4.424 | 2.758 | 2.556 | 0.803 | 1.402 |
| Qwen3.5-27B | Base    | 2048     | 4.737 | 2.949 | 2.708 | 0.806 | 1.476 |
| Qwen3.5-27B | Base    | 6144     | 4.799 | 2.946 | 2.727 | 0.809 | 1.484 |
| Qwen3.5-27B | LoRA    | 512      | 3.888 | 2.515 | 2.574 | 0.808 | 1.176 |
| Qwen3.5-27B | LoRA    | 2048     | 4.046 | 2.602 | 2.657 | 0.817 | 1.233 |
| Qwen3.5-27B | LoRA    | 6144     | 4.143 | 2.623 | 2.687 | 0.818 | 1.200 |
| Gemma4-31B  | Base    | 512      | 4.813 | 2.887 | 2.729 | 0.790 | 0.919 |
| Gemma4-31B  | Base    | 2048     | 4.894 | 2.979 | 2.848 | 0.804 | 0.994 |
| Gemma4-31B  | Base    | 6144     | 4.933 | 3.027 | 2.908 | 0.803 | 1.018 |
| Gemma4-31B  | LoRA    | 512      | 4.745 | 2.859 | 2.506 | 0.800 | 1.329 |
| Gemma4-31B  | LoRA    | 2048     | 4.890 | 2.987 | 2.623 | 0.809 | 1.420 |
| Gemma4-31B  | LoRA    | 6144     | 4.923 | 3.048 | 2.653 | 0.812 | 1.462 |
| OLMo3.1-32B | Base    | 512      | 4.118 | 2.530 | 2.397 | 0.788 | 1.081 |
| OLMo3.1-32B | Base    | 2048     | 4.236 | 2.700 | 2.509 | 0.789 | 1.196 |
| OLMo3.1-32B | Base    | 6144     | 4.274 | 2.726 | 2.484 | 0.789 | 1.330 |
| OLMo3.1-32B | LoRA    | 512      | 3.937 | 2.516 | 2.563 | 0.800 | 1.217 |
| OLMo3.1-32B | LoRA    | 2048     | 4.159 | 2.582 | 2.648 | 0.809 | 1.143 |
| OLMo3.1-32B | LoRA    | 6144     | 4.235 | 2.628 | 2.653 | 0.811 | 1.171 |

Table 17: Effect of evaluation history length for the largest QLoRA-trained models with complete cross-context evaluations. The longest condition uses a 6,000-token history limit and is labeled 6,144 to match the nominal bucket. Sem. is the average of content and stance similarity.

## **N Training Context and Duration Robustness**

We report results across 30 experimental conditions: 3 baselines (one per eval context) and 27 fine-tuned conditions (3 training contexts  $\times$  3 eval contexts  $\times$  3 epochs). Each condition is evaluated on all test examples.

### **N.1 Main Results**

Table 18 presents all 30 conditions.

### **N.2 Effect of Evaluation Context Length**

Evaluation context length has the clearest effect in this sweep. For the baseline, moving from 512 to 6,144 tokens increases believability from 3.14 to 3.53 and semantic similarity from 2.58 to 2.75. The model trained on 6,144-token histories for 15 epochs shows the same direction: believability rises from 3.19 with 512 evaluation tokens to 3.67 with 6,144, and semantic similarity rises from 2.32 to 2.48. Richer evaluation history therefore usually helps, especially when moving from short to moderate context.

### **N.3 Effect of Training Context Length**

Training on longer histories has smaller and less monotonic effects than evaluating with longer histories. At epoch 15 with 6,144 evaluation tokens, cosine similarity changes only from 0.809 after training on 512-token histories to 0.813 after training on 6,144-token histories. Style changes from 2.56 to 2.62, while believability is essentially tied at 3.70 and 3.67. In this setting, longer training context helps some reference-similarity metrics, but it does not substitute for providing more user history at inference time.

### **N.4 Effect of Training Duration**

Longer training generally improves the model trained and evaluated with 6,144-token histories, but the gains are modest. From epoch 5 to 15, cosine similarity moves from 0.809 to 0.813, content similarity from 2.09 to 2.17, and believability from 3.40 to 3.67. These later checkpoints preserve the cross-model result: training duration and context length matter, but model family and scale remain central determinants of performance.

| Train Ctx                                     | Eval Ctx | Epoch | Cosine↑      | Content↑ | Stance↑ | Style↑      | Believ.↑    | Len R. | Lev.↑        | Jacc.↑       |
|-----------------------------------------------|----------|-------|--------------|----------|---------|-------------|-------------|--------|--------------|--------------|
| <i>Baseline (untrained Qwen3-4B-Instruct)</i> |          |       |              |          |         |             |             |        |              |              |
| –                                             | 512      | –     | 0.798        | 2.20     | 2.95    | 2.41        | 3.14        | 1.44   | 0.155        | 0.132        |
| –                                             | 2048     | –     | 0.803        | 2.30     | 3.14    | 2.46        | 3.43        | 1.67   | 0.157        | 0.139        |
| –                                             | 6144     | –     | 0.805        | 2.31     | 3.19    | 2.47        | 3.53        | 1.73   | 0.160        | 0.142        |
| <i>Fine-tuned (LoRA, ctx512 training)</i>     |          |       |              |          |         |             |             |        |              |              |
| 512                                           | 512      | 5     | 0.801        | 1.97     | 2.57    | 2.39        | 3.07        | 1.21   | 0.160        | 0.129        |
| 512                                           | 2048     | 5     | 0.804        | 2.07     | 2.71    | 2.51        | 3.48        | 1.20   | 0.169        | 0.136        |
| 512                                           | 6144     | 5     | 0.805        | 2.06     | 2.68    | 2.50        | 3.49        | 1.19   | 0.180        | 0.147        |
| 512                                           | 512      | 10    | 0.804        | 2.04     | 2.65    | 2.46        | 3.18        | 1.18   | 0.169        | 0.138        |
| 512                                           | 2048     | 10    | 0.809        | 2.08     | 2.75    | 2.57        | 3.47        | 1.18   | 0.175        | 0.139        |
| 512                                           | 6144     | 10    | 0.809        | 2.14     | 2.78    | 2.54        | 3.64        | 1.18   | 0.186        | 0.152        |
| 512                                           | 512      | 15    | 0.802        | 2.03     | 2.65    | 2.49        | 3.25        | 1.19   | 0.171        | 0.135        |
| 512                                           | 2048     | 15    | 0.807        | 2.08     | 2.74    | 2.54        | 3.56        | 1.18   | 0.174        | 0.140        |
| 512                                           | 6144     | 15    | 0.809        | 2.14     | 2.81    | 2.56        | 3.70        | 1.20   | 0.186        | 0.153        |
| <i>Fine-tuned (LoRA, ctx2048 training)</i>    |          |       |              |          |         |             |             |        |              |              |
| 2048                                          | 512      | 5     | 0.801        | 1.97     | 2.56    | 2.45        | 3.06        | 1.23   | 0.159        | 0.128        |
| 2048                                          | 2048     | 5     | 0.809        | 2.04     | 2.68    | 2.53        | 3.36        | 1.25   | 0.165        | 0.134        |
| 2048                                          | 6144     | 5     | 0.811        | 2.12     | 2.74    | 2.56        | 3.49        | 1.24   | 0.181        | 0.151        |
| 2048                                          | 512      | 10    | 0.800        | 1.99     | 2.60    | 2.48        | 3.23        | 1.17   | 0.160        | 0.127        |
| 2048                                          | 2048     | 10    | 0.812        | 2.12     | 2.74    | 2.58        | 3.41        | 1.25   | 0.174        | 0.142        |
| 2048                                          | 6144     | 10    | 0.811        | 2.14     | 2.77    | 2.56        | 3.50        | 1.24   | 0.182        | 0.153        |
| 2048                                          | 512      | 15    | 0.800        | 2.01     | 2.60    | 2.43        | 3.22        | 1.16   | 0.164        | 0.130        |
| 2048                                          | 2048     | 15    | 0.807        | 2.09     | 2.75    | 2.56        | 3.49        | 1.19   | 0.175        | 0.142        |
| 2048                                          | 6144     | 15    | 0.811        | 2.11     | 2.73    | 2.58        | 3.62        | 1.18   | 0.187        | 0.153        |
| <i>Fine-tuned (LoRA, ctx6144 training)</i>    |          |       |              |          |         |             |             |        |              |              |
| 6144                                          | 512      | 5     | 0.799        | 1.95     | 2.52    | 2.38        | 3.08        | 1.27   | 0.158        | 0.126        |
| 6144                                          | 2048     | 5     | 0.807        | 2.05     | 2.68    | 2.50        | 3.36        | 1.29   | 0.171        | 0.139        |
| 6144                                          | 6144     | 5     | 0.809        | 2.09     | 2.70    | 2.52        | 3.40        | 1.26   | 0.183        | 0.149        |
| 6144                                          | 512      | 10    | 0.800        | 2.00     | 2.59    | 2.46        | 3.27        | 1.17   | 0.165        | 0.131        |
| 6144                                          | 2048     | 10    | 0.808        | 2.09     | 2.75    | 2.59        | 3.55        | 1.17   | 0.176        | 0.142        |
| 6144                                          | 6144     | 10    | 0.812        | 2.13     | 2.76    | 2.62        | 3.61        | 1.18   | 0.188        | 0.155        |
| 6144                                          | 512      | 15    | 0.798        | 2.02     | 2.62    | 2.44        | 3.19        | 1.20   | 0.165        | 0.130        |
| 6144                                          | 2048     | 15    | 0.810        | 2.09     | 2.73    | 2.57        | 3.56        | 1.16   | 0.178        | 0.144        |
| 6144                                          | 6144     | 15    | <b>0.813</b> | 2.17     | 2.80    | <b>2.62</b> | <b>3.67</b> | 1.20   | <b>0.191</b> | <b>0.157</b> |

Table 18: Full results across all conditions. **Bold** marks the best fine-tuned result per metric. Baselines receive the same history in-context without parameter updates. LLM judge scores are on a 1–5 scale (higher is better). Len R. = length ratio of predicted to gold reply (1.0 is ideal).