MUSIC: MUlti-Step Instruction Contrast for Multi-Turn Reward Models

Anonymous ACL submission

Abstract

Evaluating the quality of multi-turn conversations is crucial for developing capable Large Language Models (LLMs), yet remains a significant challenge, often requiring costly human evaluation. Multi-turn reward models (RMs) offer a scalable alternative and can provide valuable signals for guiding LLM training. While recent work has advanced multi-turn *training* techniques, effective automated *evaluation* specifically for multi-turn interactions lags behind. We observe that standard preference datasets, typically contrasting responses based only on the final conversational turn, provide insufficient signal to capture the nuances of multi-turn interactions. Instead, we find that incorporating contrasts spanning *multiple* turns is critical for building robust multi-turn RMs. Motivated by this finding, we propose **MU**lti-**S**tep **I**nstruction **C**ontrast (MUSIC), an unsupervised data augmentation strategy that synthesizes contrastive conversation pairs exhibiting differences across multiple turns. Leveraging MUSIC on the Skywork preference dataset, we train a multi-turn RM based on the Gemma-2-9B-Instruct model. Empirical results demonstrate that our MUSIC-augmented RM outperforms baseline methods, achieving higher alignment with judgments from advanced proprietary LLM judges on multi-turn conversations, crucially, without compromising performance on standard single-turn RM benchmarks.

1 Introduction

The ability of Large Language Models (LLMs) to engage in coherent, multi-turn conversations is a hallmark of advanced AI systems (Turing, 1950). While recent LLMs demonstrate remarkable proficiency in single-turn instruction following and short dialogues (Ouyang et al., 2022; Adler et al., 2024; Team et al., 2023), extending this capability to complex, long-horizon interactions remains a critical frontier (Zheng et al., 2023; Abdulhai et al., 2023; He et al., 2024; Deshpande et al., 2025). Significant effort has focused on developing Reinforcement Learning from Human Feedback (RLHF) techniques tailored for multi-turn dynamics (Zhou et al., 2024; Gao et al., 2024; Shi et al., 2024; Shani et al., 2024; He et al., 2025; Abdulhai et al., 2025; Jiang et al., 2025), aiming to improve conversational performance beyond standard single-turn RLHF methods.

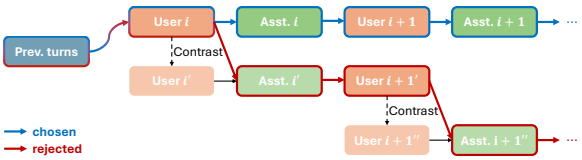

Figure 1: Overview of the MUSIC data augmentation procedure. Given seed contexts from existing datasets, we generate multi-turn rollouts where LLM simulators generate contrastive pairs, and use a contrastive instruction prompt to induce quality degradation in the rejected branch. The augmented preference pairs are used to train a multi-turn reward model along with the original dataset. Black arrows represent ephemeral changes that are provided to the assistant once, but not persisted. For each augmented pair, the chosen example consists of turns with blue borders, while the rejected example consists of turns with red borders.

Despite advances in multi-turn *training*, robust automated *evaluation* of these interactions is a persistent challenge. High-quality, model-based evaluators, or specifically reward models (RMs), are crucial, serving not only as direct performance metrics but also providing signals during training and inference (Lambert et al., 2024; Malik et al., 2025). However, evaluating multi-turn conversations is fundamentally more complex than single-turn assessment. It requires judging not only the response quality at each turn but also inter-turn properties like coherence, consistency, and effective use of conversational history (Deshpande et al., 2025; He et al., 2024). Therefore, training powerful multi-turn RMs typically necessitates large volumes of

high-quality preference data reflecting these nuances (Wang et al., 2024b,c; Liu et al., 2024).

Acquiring such data via human annotation is prohibitively expensive. Comparing two lengthy conversations, potentially differing subtly across multiple turns, is significantly more demanding and time-consuming than annotating single-turn preferences (Deshpande et al., 2025). As a result, widely used preference datasets (Bai et al., 2022; Ganguli et al., 2022; Ethayarajh et al., 2022; Cui et al., 2023; Liu et al., 2024) often contain predominantly single-turn pairs or multi-turn pairs where the difference is confined to the final turn. While practical for efficient data collection, this data characteristic may limit the ability of RMs trained on them to capture holistic conversational quality. This motivates our central research question:

Can we develop a scalable approach to synthesize contrastive data spanning multiple turns to train more effective multi-turn RMs?

To address this data gap, we propose **MU**lti-**S**tep **I**nstruction **C**ontrast (MUSIC), an unsupervised data augmentation strategy designed to generate contrastive conversation pairs with meaningful quality differences distributed across multiple turns, without human annotation. By introducing controlled variations of instructions during the generation process, one conversation in the pair is constructed to be qualitatively better (e.g., more consistent, exhibiting better instruction following) than the other *across multiple turns*. This creates contrastive data specifically highlighting multi-turn phenomena where the quality distinction is woven throughout the conversation. MUSIC can be readily applied to augment existing preference datasets, enriching them with multi-turn contrast signals.

We demonstrate the efficacy of MUSIC by applying it to the Skywork preference dataset (Liu et al., 2024) and fine-tuning a Gemma-2-9B-Instruct model on this augmented data. Our experiments show that the resulting MUSIC-augmented RM maintains strong performance on standard single-turn benchmarks like RewardBench (Lambert et al., 2024). More importantly, compared to baseline models trained without MUSIC, our RM exhibits higher agreement with judgments from the advanced Gemini 1.5 Pro model when assessing the quality of multi-turn conversations.

Our contributions are threefold:

1. We identify a critical limitation in standard preference datasets for training multi-turn RMs: the predominant focus on final-turn contrasts, which hinders the learning of holistic conversational quality assessment.
2. We propose MUSIC, a scalable, unsupervised method to synthesize contrastive conversation pairs with meaningful quality differences spanning multiple turns, directly addressing the identified data gap.
3. We demonstrate empirically that RMs trained with MUSIC achieve improved alignment with advanced LLM judges on multi-turn tasks, without sacrificing performance on single-turn benchmarks, validating the effectiveness of our approach.

2 Related Work

Due to space constraints, we provide a brief overview of related work here and defer to Appendix A for a more comprehensive discussion.

Reward Modeling in Multi-Turn Contexts. Aligning LLMs typically relies on reward models (RMs) to guide policy optimization (Ouyang et al., 2022; Bai et al., 2022). Recent literature distinguishes between *Outcome RMs* (ORMs), which score entire generations (Liu et al., 2024; Wang et al., 2024c), and *Process RMs* (PRMs), which evaluate intermediate steps (Lightman et al., 2023; Wang et al., 2023a). While PRMs offer dense supervision, they require expensive fine-grained annotations. In multi-turn dialogue, where long-term credit assignment is challenging (Abdulhai et al., 2023), standard ORMs trained on preference data where contrasts are localized often struggle to capture implicit quality signals like coherence and consistency. Unlike PRMs or existing multi-turn benchmarks that rely on static rubrics (He et al., 2024, 2025), MUSIC enhances standard ORMs by synthesizing preference pairs with distributed quality differences. This allows the RM to learn holistic multi-turn properties, providing robust signals necessary for advanced multi-turn RL algorithms (Zhou et al., 2024; Jiang et al., 2025).

Synthetic Data for Alignment. The scarcity of high-quality human data has driven the use of LLMs to generate synthetic data for supervised fine-tuning and feedback (Wang et al., 2023b; Dubois et al., 2023; Chen et al., 2024). While recent works extend this paradigm to multi-turn conversations (Wu et al., 2025a; Yin et al., 2025a), they

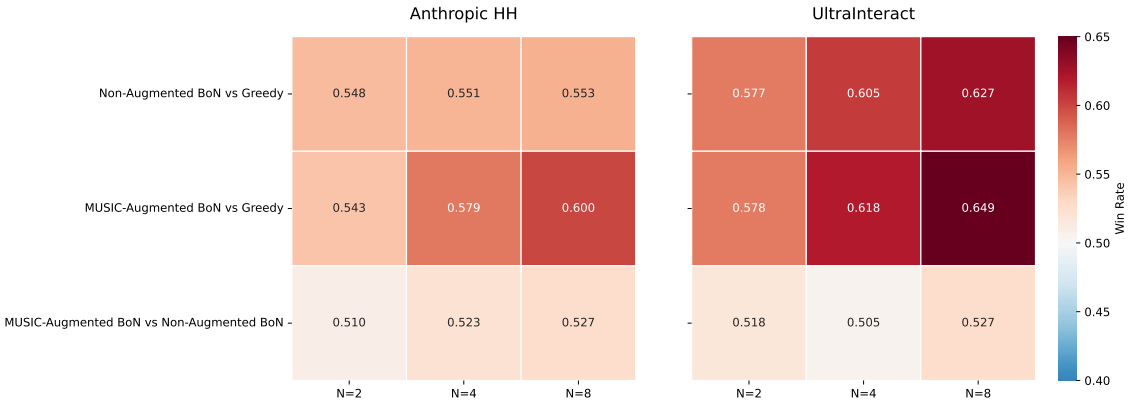

Figure 2: Winrates comparing conversations generated via Best-of-N ($N \in \{2, 4, 8\}$) guided by the MUSIC-Augmented RM versus the Baseline (non-augmented) RM, evaluated by Gemini 1.5 Pro on subsets of Anthropic HH and UltraInteract. Comparisons against greedy decoding are also shown.

primarily focus on generating data for policy training (SFT or RL). MUSIC distinguishes itself by targeting the *reward modeling* stage: it automates the creation of *contrastive preference pairs* via controlled noise injection. This acts as a targeted data augmentation technique, enabling RMs to better discriminate between coherent and incoherent multi-turn trajectories.

3 Method

We introduce **MU**lti-**S**tep **I**nstruction **C**ontrast (MUSIC), a scalable, unsupervised method for synthesizing contrastive conversation pairs that exhibit meaningful quality differences across multiple turns. This synthesized data is designed to augment existing preference datasets, enabling the training of more effective multi-turn RMs. The core process involves three stages:

Initialization: We sample conversational prefixes (seed contexts) from an existing multi-turn dataset to initiate the augmentation process.

Multi-turn Rollouts with MUSIC: Starting from each seed context, we employ LLM-based user and assistant simulators to generate paired conversations. Crucially, at each turn, a contrastive instruction prompt guides the assistant simulator to produce a lower-quality response for one conversation in the pair.

Multi-turn RM Training: The conversation pairs generated by MUSIC are combined with the original preference data. A multi-turn RM is then trained on this augmented dataset using standard preference learning techniques.

Due to space limits, we defer a detailed description of our method to Appendix B.

4 Experiments

Our experiments are designed to investigate the efficacy of MUSIC by addressing the following research questions: **(a)** Does MUSIC improve the effectiveness of RMs for assessing multi-turn conversations? **(b)** Does augmenting training data with MUSIC negatively impact the RM's performance on standard single-turn RM benchmarks?

To answer **(a)**, we evaluate the performance of a MUSIC-augmented RM against a baseline RM (trained without MUSIC) in a multi-turn Best-of-N (BoN) inference task. This task requires the RM to iteratively select the best response from N candidates generated by an assistant LLM at each turn of a conversation. The quality of the resulting multi-turn conversations serves as a proxy for the RM's effectiveness. To answer **(b)**, we evaluate both RMs on RewardBench (Lambert et al., 2024), a standard benchmark primarily focused on single-turn evaluation capabilities.

4.1 Experimental Setup

We introduce our experimental setup as follows, and defer additional details to Appendix C.

Data. We train RMs on Skywork-Reward-Preference-80K-v0.2 (Dorka, 2024; Shiwen et al., 2024; Liu et al., 2024), which is dominated by single-turn pairs and multi-turn pairs differing only in the final turn, making it well-suited for MUSIC augmentation. We filter to dialogues with ≤ 5 turns and remove samples exceeding 2048 tokens. MUSIC uses Gemini 1.5 Pro to simulate user/assistant rollouts with T=5. The resulting training data contains ∼73K original preference pairs and ∼31K MUSIC pairs.

Table 1: RewardBench accuracy (%) results. We compare the RM trained on the original Skywork dataset and the RM trained on the MUSIC-augmented dataset. Both use Gemma-2-9B-Instruct as the base model.

Model	Overall	Chat	Chat Hard	Safety	Reasoning
Gemma-2-9B-Instruct w/ Skywork	85.7	**91.9**	83.8	88.4	78.6
Gemma-2-9B-Instruct w/ Skywork + MUSIC	**87.2**	91.6	**85.1**	**89.7**	**82.5**

Models and training. We fine-tune Gemma-2-9B-Instruct (Team et al., 2024) with a scalar reward head, training (i) a baseline RM on $\mathcal{D}$ and (ii) a MUSIC-augmented RM on $\mathcal{D}_{\text{aug}}$. Both use Bradley–Terry loss with AdamW (Loshchilov and Hutter, 2017) optimizer.

Evaluation. We evaluate multi-turn Best-of-N inference, where at each of $H{=}3$ turns the assistant samples $N \in \{2, 4, 8\}$ candidates and the RM selects the continuation, and the final conversations are judged by Gemini 1.5 Pro. We also report single-turn performance on RewardBench (Lambert et al., 2024) using the standard pairwise-accuracy protocol.

4.2 Results on Multi-Turn Best-of-N Inference

Figure 2 presents the winrates from the multi-turn BoN evaluation. We compare conversations generated using BoN guided by the MUSIC-Augmented RM against those guided by the Baseline RM, as judged by Gemini 1.5 Pro. For reference, we also include comparisons against greedy decoding from the assistant LLM.

Across both the Anthropic HH and UltraInteract initial prompts, the results consistently demonstrate that conversations guided by the MUSIC-Augmented RM are preferred over those guided by the Baseline RM. Furthermore, the performance gap generally widens as N increases, indicating that the MUSIC-Augmented RM effectively leverages the stronger candidate pool provided by larger N. Both BoN methods outperform the greedy baseline substantially. This provides strong evidence for research question (a): **MUSIC successfully enhances the RM's ability to identify and promote higher-quality multi-turn interactions**, leading to demonstrably better conversational outputs as judged by an advanced LLM.

4.3 Results on RewardBench

Table 1 shows the performance of the Baseline and MUSIC-Augmented RMs on RewardBench. Addressing research question (b), we observe that **augmenting the training data with MUSIC does not sacrifice single-turn evaluation performance**. In fact, the MUSIC-Augmented RM achieves slightly better or comparable accuracy across the Chat, Chat Hard, and Safety categories.

Surprisingly, we observe a notable improvement (+3.9%) in the Reasoning category for the MUSIC-Augmented RM. While MUSIC synthesizes multi-turn conversational data and is not explicitly designed for single-turn reasoning tasks, this suggests a potential positive transfer. We hypothesize that exposure to coherent, logically structured multi-turn dialogues during training may implicitly enhance the RM's ability to assess reasoning steps, even when presented in single turns. Overall, these results indicate that **MUSIC not only improves multi-turn evaluation capabilities but does so without compromising, and potentially even slightly enhancing, performance on standard single-turn benchmarks**.

5 Conclusion

In this work, we addressed the challenge of evaluating multi-turn conversations by introducing **MU**lti-**S**tep **I**nstruction **C**ontrast (MUSIC), a scalable, unsupervised data augmentation technique. MUSIC synthesizes contrastive conversation pairs where quality differences are intentionally distributed across multiple turns, enriching standard preference datasets that often focus on final-turn contrasts. We demonstrated that training a multi-turn RM on a MUSIC-augmented dataset leads to improved performance in guiding multi-turn interactions, as measured by alignment with judgments from an advanced LLM judge in a Best-of-N setting. Crucially, these gains in multi-turn evaluation capability were achieved without compromising, and potentially even slightly enhancing, performance on standard single-turn benchmarks like RewardBench. Our results validate MUSIC as an effective strategy for training more robust multi-turn RMs, mitigating the need for expensive human annotation of complex conversational preferences.

Limitations

While promising, our work has several limitations that suggest avenues for future research.

Reliance on LLM Simulators and Judges: Both the MUSIC data generation process (using M_u and M_a) and the primary multi-turn evaluation (BoN judged by Gemini 1.5 Pro) rely heavily on LLMs. While practical and scalable, these models may introduce their own biases or fail to capture the full spectrum of human conversational nuances and preferences. Future work could explore incorporating real human interactions or judgments, potentially through targeted human-in-the-loop refinement or evaluation on human-annotated multi-turn benchmarks, to further validate and potentially improve the approach.

Conversation Length and Model Scale: Our experiments were constrained by computational resources and model context windows, limiting MUSIC rollouts to $T = 5$ turns and BoN evaluation to $H = 3$ turns. The effectiveness of MUSIC for significantly longer conversations remains to be explored. Scaling the approach to larger base models with longer context windows is a natural next step, potentially unlocking benefits for evaluating more complex, extended dialogues.

Addressing these limitations represents promising directions for advancing automated evaluation of complex, multi-turn LLM interactions.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*.

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*.

Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, and 1 others. 2024. Gpt-4 technical report, 2024. *URL https://arxiv. org/abs/2303.08774*, 2:6.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *arXiv preprint arXiv:2409.10164*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Zhaolin Gao, Wenhao Zhan, Jonathan D Chang, Gokul Swamy, Kianté Brantley, Jason D Lee, and Wen Sun. 2024. Regressing the relative future: Efficient policy optimization for multi-turn rlhf. *arXiv preprint arXiv:2410.04612*.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, and 1 others. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Selina Peng, Beibin Li, and 1 others. 2025. Rubric-based benchmarking and reinforcement learning for advancing llm instruction following. *arXiv preprint arXiv:2511.10507*.

Daniel R Jiang, Jalaj Bhandari, Yukai Yang, Rémi Munos, and Tyler Lu. 2025. Aligning llms toward multi-turn conversational outcomes using iterative ppo. *arXiv preprint arXiv:2511.21638*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1192–1202.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, and 1 others. 2024. Multi-turn reinforcement learning from preference human feedback. *arXiv preprint arXiv:2405.14655*.

Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. 2024. Direct multi-turn preference optimization for language agents. *arXiv preprint arXiv:2406.14868*.

Tu Shiwen, Zhao Liang, Chris Yuhao Liu, Liang Zeng, and Yang Liu. 2024. Skywork critic model series. https://huggingface.co/Skywork.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *ACL*.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2023a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024b. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024c. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Jiangxu Wu, Cong Wang, TianHuang Su, Jun Yang, Haozhi Lin, Chao Zhang, Ming Peng, Kai Shi, SongPan Yang, BinQing Pan, and 1 others. 2025a. Instruct: A review-driven multi-turn conversations generation method for large language models. *arXiv preprint arXiv:2505.11010*.

Shujin Wu, Yi R Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025b. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

Shangjian Yin, Zhepei Wei, Xinyu Zhu, Wei-Lin Chen, and Yu Meng. 2025a. Aligning large language models via fully self-synthetic data. *arXiv preprint arXiv:2510.06652*.

Zhangyue Yin, Qiushi Sun, Zhiyuan Zeng, Qinyuan Cheng, Xipeng Qiu, and Xuan-Jing Huang. 2025b. Dynamic and generalizable process reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4203–4233.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024a. Advancing llm reasoning generalists with preference trees. *Preprint*, arXiv:2404.02078.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024b. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.

Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, and 1 others. 2025. Versaprm: Multi-domain process reward model via synthetic reasoning data. *arXiv preprint arXiv:2502.06737*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Related Work

Preference Learning and Reward Modeling. Aligning LLMs with human values has evolved significantly since the foundational frameworks of Reinforcement Learning from Human Feedback (RLHF) were established (Christiano et al., 2017; Ziegler et al., 2019). The standard pipeline relies on learning a reward model (RM) from human preferences to guide policy optimization (Ouyang et al., 2022; Bai et al., 2022). While alternatives like Direct Preference Optimization (DPO) (Rafailov et al., 2023) bypass explicit reward modeling, RMs remain essential for scalable oversight, rejection sampling, and guiding search (Lambert et al., 2024), especially in domains without verifiable rewards. Recent literature on RMs has bifurcated into two distinct streams:

- **Outcome Reward Models (ORMs):** These models typically assign a single scalar score to an entire LLM generation (e.g., a full response or conversational turn) based on its overall quality (Liu et al., 2024; Wang et al., 2024c; Cobbe et al., 2021; Wang et al., 2024a). They are widely used for general instruction following, dialogue, and reasoning tasks.

- **Process Reward Models (PRMs):** These provide denser supervision by evaluating intermediate steps within a generation process, such as individual reasoning steps in mathematical proofs or lines of code (Uesato et al., 2022; Lightman et al., 2023; Wang et al., 2023a), and recently extending to more general domains (Zeng et al., 2025; Yin et al., 2025b). However, PRMs require more fine-grained annotations and thus are more expensive to train compared to ORMs.

Our work focuses on enhancing ORMs for open-ended *multi-turn conversations*, where "steps" are conversational turns and the quality signal is often implicit and distributed rather than discrete and verifiable. While standard ORM training often relies on preference data where contrasts are localized (e.g., single-turn differences or final-turn edits in dialogues), MUSIC acts as a data augmentation technique. It synthesizes preference pairs where the quality difference is intentionally distributed across multiple turns. By training standard ORM architectures on MUSIC-augmented data, we aim to improve their ability to capture holistic multi-turn properties like coherence and consistency, which are often underspecified by conventional preference datasets and distinct from the step-level focus of PRMs.

Multi-Turn Alignment. Extending alignment to multi-turn interactions introduces significant complexity due to long-term credit assignment challenges (Abdulhai et al., 2023). Early dialogue systems often use handcrafted reward functions based on heuristics (Li et al., 2016) for RL on small-scale models, while more recent approaches investigate RL techniques on LLM tailored for multi-turn alignment, including but not limited to hierarchical RL (Zhou et al., 2024), value-based Jiang et al. (2025) and self-play or multi-agent Shani et al. (2024); Wu et al. (2025b) methods. However, these advanced policy optimization methods depend critically on robust reward signals. While existing multi-turn benchmarks (He et al., 2024; Deshpande et al., 2025; He et al., 2025) leverage human annotations or rubric-based methods for evaluation, such efforts are often costly and not scalable for training. Our work complements this line of work by enhancing the underlying reward models through MUSIC, thereby improving the overall multi-turn alignment process.

Synthetic Data for Alignment. The scarcity of high-quality human annotations has driven a shift toward synthetic data generation. Recent work demonstrates that LLMs could generate their own fine-tuning data (Wang et al., 2023b; Dubois et al., 2023) and provide feedback signals for improvement (Chen et al., 2024; Yuan et al., 2024b). This paradigm is also used to generate multi-turn conversations for LLM training more recently (Wu et al., 2025a; Yin et al., 2025a). Unlike these methods, which primarily focus on generating data for SFT or RL, MUSIC focuses specifically on synthesizing *contrastive preference pairs* to train a multi-turn RM. We automate the creation of chosen and rejected trajectories by injecting controlled noise, thereby providing the necessary discriminative signals.

B Algorithm Details

In this section, we first review the preliminaries for training model-based RMs, and then describe each stage of our pipeline in detail.

B.1 Preliminaries

We focus on ORMs, where the model R_θ maps a conversation (or parts thereof) to a scalar score (Liu et al., 2024; Wang et al., 2024c). Training typically involves maximizing the log-likelihood of observing human preferences under the Bradley-Terry (BT) model (Bradley and Terry, 1952):

$$\mathcal{L}(\theta, \mathcal{D}) = \mathbb{E}_{C_{\text{chosen}}, C_{\text{rejected}} \sim \mathcal{D}} \log \sigma \left(R_\theta(C_{\text{chosen}}) - R_\theta(C_{\text{rejected}}) \right) \quad (1)$$

where $\mathcal{D}$ is a preference dataset of conversation pairs $(C_{\text{chosen}}, C_{\text{rejected}})$, and σ is the sigmoid function, respectively. In practice, R_θ is often implemented by fine-tuning a pre-trained or instruction-tuned LLM, adding a linear layer to map a representation (e.g., the last-layer hidden state of the final token) to the scalar reward score.

B.2 Initialization

We assume access to an existing multi-turn preference dataset $\mathcal{D} = \{(C^{(i)}_{\text{chosen}}, C^{(i)}_{\text{rejected}})\}_{i=1}^{N}$. As noted earlier, such datasets (Bai et al., 2022; Ganguli et al., 2022; Liu et al., 2024) often contain pairs differing only in the final turn, providing limited signal for multi-turn phenomena. However, the initial turns often represent valid, human-generated conversational trajectories. We leverage this by sampling seed contexts from $\mathcal{D}$. Specifically, for conversation $C^{(i)}$ in the dataset of H turns, we sample a turn index $h \sim U(1, H)$ uniformly at random and extract the first h turns as the seed context $C_{\text{prefix}} = C^{(i)}_{1:h}$. This approach balances the reuse of high-quality human-curated conversational prefixes with the generation of novel multi-turn contrasts via MUSIC.

B.3 Multi-turn Rollouts with MUSIC

Given a set of seed contexts, we apply the MUSIC algorithm (Algorithm 1) to generate contrastive conversation pairs $\mathcal{D}_{\text{MUSIC}}$. This process simulates multi-turn interactions using LLMs as proxies for both the user (M_u) and the assistant (M_a), inspired by work on generative agents (Park et al., 2023, 2024).

The core idea of MUSIC is to introduce controlled quality degradation in one branch of the simulated conversation pair at each turn. This is achieved via the instruction contrast prompt, $\texttt{Contrast}(\cdot)$. For the *chosen* conversation path C_{chosen}, the simulated assistant M_a responds directly to the simulated user's utterance u_t^{chosen}. For the *rejected* path C_{rejected}, however, the user's utterance u_t^{rejected} is first transformed by $\text{Contrast}(\cdot)$ into a modified instruction, which prompts M_a to generate a response a_t^{rejected} that is intentionally suboptimal relative to the original user utterance u_t^{rejected} (e.g., less helpful, inconsistent with previous turns, or failing to follow a specific constraint). As shown in Figure 1, the instruction contrast prompt implicitly guides the assistant to generate responses through ephemeral modifications, ensuring the rejected trajectory remains coherent yet qualitatively inferior to its chosen counterpart. The design details of $\texttt{Contrast}(\cdot)$ are provided in Appendix D.3, drawing inspiration from (Wang et al., 2024b).

By repeating this process for T turns, MUSIC generates paired conversations $(C_{\text{chosen}}, C_{\text{rejected}})$ where C_{chosen} is superior by construction, and the quality difference is distributed across multiple turns rather than being localized. This yields preference data specifically designed to train RMs sensitive to multi-turn conversational dynamics.

Algorithm 1 **MU**lti-**S**tep **I**nstruction **C**ontrast (MUSIC) Data Generation

Require: Seed conversation context C_{prefix}, LLM user simulator M_u, LLM assistant simulator M_a, max simulation turns T, instruction contrast prompt $\texttt{Contrast}(\cdot)$

Initialize $C_{\text{chosen}} \leftarrow C_{\text{prefix}}$, $C_{\text{rejected}} \leftarrow C_{\text{prefix}}$

for $t = 1$ **to** T **do**

Generate next user utterance: $u_t^{\text{chosen}} \leftarrow M_u(C_{\text{chosen}})$, $u_t^{\text{rejected}} \leftarrow M_u(C_{\text{rejected}})$

Generate chosen assistant response: $a_t^{\text{chosen}} \leftarrow M_a(C_{\text{chosen}} \oplus u_t^{\text{chosen}})$

Generate rejected assistant response: $a_t^{\text{rejected}} \leftarrow M_a(C_{\text{rejected}} \oplus \texttt{Contrast}(u_t^{\text{rejected}}))$

Append turn to the context:

$$C_{\text{chosen}} \leftarrow C_{\text{chosen}} \oplus (u_t^{\text{chosen}}, a_t^{\text{chosen}}),$$
$$C_{\text{rejected}} \leftarrow C_{\text{rejected}} \oplus (u_t^{\text{rejected}}, a_t^{\text{rejected}})$$

end for

return $(C_{\text{chosen}}, C_{\text{rejected}})$

B.4 Multi-turn RM Training

After generating the MUSIC dataset $\mathcal{D}_{\text{MUSIC}}$, we create the final augmented training dataset $\mathcal{D}_{\text{aug}} = \mathcal{D} \cup \mathcal{D}_{\text{MUSIC}}$. We then train our multi-turn RM R_θ on $\mathcal{D}_{\text{aug}}$ by optimizing the BT loss objective in Equation 1. We train for a small number of epochs (e.g., less than two) to mitigate potential overfitting to the combined dataset. The resulting RM R_θ is expected to have improved sensitivity to multi-turn conversational properties due to its exposure to the contrastive examples synthesized by MUSIC.

C Experiment Details

Dataset Construction. We use Skywork-Reward-Preference-80K-v0.2 as the RM training dataset as it is used to train several state-of-the-art RMs (Dorka, 2024; Shiwen et al., 2024; Liu et al., 2024). This dataset is representative of standard preference data, containing mostly single-turn pairs and multi-turn pairs differing only in the final turn, making it a suitable candidate for augmentation with MUSIC. Specifically, we filter the dataset to include only dialogues with at

most five turns and uniformly sample the seed contexts as described in Section B.2. For MUSIC augmentation, we use Gemini 1.5 Pro as both user and assistant simulators with distinct prompts (see Appendix D.1 and D.2), and set the maximum simulation turns $T = 5$. Both $\mathcal{D}$ and $\mathcal{D}_{\text{MUSIC}}$ are preprocessed by filtering out conversations exceeding 2048 tokens (the maximum sequence length for training). Our final datasets consist of approximately 73k pairs from the original Skywork-Reward-Preference-80K-v0.2 dataset and 31k pairs from the MUSIC augmentation.

Training Details. We fine-tune Gemma-2-9B-Instruct (Team et al., 2024) to create our RMs. A linear layer is added on top of the transformer's final hidden state output to produce a scalar reward score. We train two main models:

1. **Baseline RM:** Trained on $\mathcal{D}$.
2. **MUSIC-Augmented RM:** Trained on $\mathcal{D}_{\text{aug}}$.

Both models are trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-6}, a global batch size of 64, and a maximum sequence length of 2048. We use a cosine learning rate decay schedule and train for 2500 steps to minimize the Bradley-Terry loss (Equation 1).

Evaluation Details. We compare the Baseline RM and the MUSIC-Augmented RM on the following tasks:

1. **Multi-Turn Best-of-N (BoN) Inference:** Best-of-N is an effective approach to leverage single-turn RMs to improve LLMs' single-turn capability (Brown et al., 2024; Wu et al., 2024; Snell et al., 2024). This task assesses the RM's ability to guide an LLM assistant towards generating higher-quality multi-turn conversations. We simulate interactions between a user (Gemini 1.5 Pro) and an assistant (Gemma-2-9B-Instruct). Conversations are initiated using 1000 prompts sampled from subsets of Anthropic HH (Bai et al., 2022) and UltraInteract (Yuan et al., 2024a), following (Gao et al., 2024). At each of $H = 3$ turns, the assistant generates $N \in \{2, 4, 8\}$ candidate responses at a fixed temperature. The RM being tested selects the response with the highest score, which is then used to continue the conversation. The number of turns $H = 3$ was chosen to accommodate the 2048 context length of the RMs and assistant. After H turns, the quality of the full conversation generated using the MUSIC-Augmented RM is compared against the conversation generated using the Baseline RM. We use Gemini 1.5 Pro as an LLM judge, prompting it to select the better conversation based on criteria adapted from (Zheng et al., 2023) (prompt in Appendix D.4). To mitigate positional bias, each pair of conversations is evaluated twice with the order swapped, and we report the average winrate. We also compare against greedy decoding from the assistant as a reference.

2. **RewardBench:** To assess single-turn performance, we evaluate both RMs on RewardBench (Lambert et al., 2024). Following the standard protocol, we report pairwise accuracy across its four main categories (Chat, Chat Hard, Safety, Reasoning) and the overall average accuracy.

D Prompt Design

D.1 User Simulator Prompt

The prompt template for the user simulator is adapted from (Gao et al., 2024; Dubois et al., 2024; Rafailov et al., 2023):

> Below is a dialogue between the user and the assistant. Pretend you are the user in this conversation. What question would you ask next?
>
> {{previous turns}}
>
> ### Instructions:
> FIRST, provide a justification of the question you want to ask.
> SECOND, on a new line, state only the question.
> Your response should use the format:
> Justification:
> Question:

D.2 Assistant Simulator Prompt

For the assistant LLM, we directly follow the prompt template provided in (Team et al., 2024):

<start_of_turn>user
{{1st turn instruction}}<end_of_turn>
<start_of_turn>model
{{1st turn response}}<end_of_turn>
<start_of_turn>user
{{2nd turn instruction}}<end_of_turn>
<start_of_turn>model
{{2nd turn response}}<end_of_turn>
...
<start_of_turn>user
{{last turn instruction}}<end_of_turn>
<start_of_turn>model

D.3 Instruction Contrast Prompt

The instruction contrast prompt is the core to synthesize turn-level differences in MUSIC. Inspired by (Wang et al., 2024b), we directly encode the instruction contrast prompt into the prompt for the assistant LLM to generate the rejected conversations in the preference pairs:

Below is a dialogue between the user and the assistant. Pretend you are the assistant in this conversation.

{{previous turns}}

Instructions:
FIRST, generate a modified instruction that is highly relevant but not semantically identical to the instruction above from the user in the last turn.
SECOND, on a new line, generate a high-quality answer which is a good response to the modified instruction but not a good response to the original user question.
Your response should use the format:
Modified Instruction:
Answer:

D.4 Evaluator Prompt

We adapt the single-turn evaluation prompt from (Zheng et al., 2023) to evaluate multi-turn conversations:

Please act as an impartial judge and evaluate the quality of the conversation between the user and two AI assistants displayed below. You should choose the assistant that follows the user's instructions and answers the user's questions better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two conversations and provide a short explanation. Avoid any position biases and ensure that the order in which the conversations were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your evaluation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better.

[The Start of Assistant A's Conversation]
{{conversation A}}
[The End of Assistant A's Conversation]

[The Start of Assistant B's Conversation]
{{conversation B}}
[The End of Assistant B's Conversation]