# Aligning Recommendation Explanations to User Preferences Using LLMs Fine-Tuned by Reinforcement Learning with AI Feedback

Julien Albert[1][0000−0001−6279−5601], Martin Balfroid[1][0000−0002−1318−1184], Lluc Bono Rosselló[2][0000−0003−1999−6680], Arunav Das[3][0009−0008−9989−1718], Lucile Dierckx[4][0000−0003−2855−1042], and Yanni Sun[5][0000−0002−2377−1137]

[1] NaDI, UNamur, Namur, Belgium, {julien.albert,martin.balfroid}@unamur.be
[2] IRIDIA, ULB, Bruxelles, Belgium, lluc.bono.rossello@ulb.be
[3] LDC, KCL, London, UK, yanni.sun@kcl.ac.uk
[4] ICTEAM, UCLouvain, Louvain-la-Neuve, Belgium, lucile.dierckx@uclouvain.be
[5] Department of Informatics, KCL, London, UK, arunav.das@kcl.ac.uk

## 1 Approach

**Motivation.** Large Language Models (LLMs) can generate user-friendly explanations from graph-based recommendation data; yet those are perceived as only slightly more *satisfying* than template-based approaches [1]. To improve user satisfaction [2] without extensive human evaluation, we explore reinforcement learning with AI feedback (RLAIF) [3].

**Pipeline.** We use the MovieLens-32M dataset[6], selecting the one million most recent ratings from active users for computation efficiency. Additional metadata was retrieved from IMDb and enriched with tag similarity information. For recommendation, we employ an item-based KNN model, providing intrinsic explanations from the ten nearest neighbors. To provide a richer explanatory context, we construct a knowledge graph that links users, movies, associated metadata, and similar tags. Post-hoc explanations are then generated using Sorted Explanation Paths [4]. This intrinsic explanation and post-explanation are provided to an LLM tasked with generating a textual explanation.

**LLM Fine-tuning.** We aim to compare four models to assess whether odds ratio preference optimization (ORPO) [5] provides added value beyond standard supervised fine-tuning (SFT). Two of which serve as low-end (Qwen3 0.6B) and high-end (Gemini 2.5 Flash) baselines. The other two are improved versions of Qwen3 0.6B: one with SFT using synthetic data generated with Gemini 2.5 Flash, and the other with ORPO using a preference dataset (comprising preferred/rejected pairs of explanations). For the synthetic dataset, we generated 1,250 recommendations, each accompanied by both intrinsic and post-hoc explanations as described above. For the preference dataset, we generated two alternative explanations, for each recommendation, from Gemini 2.5 Flash (the teacher model) to enable preference learning. We constructed this using

[6] https://grouplens.org/datasets/movielens/32m/

AI feedback, with GPT-4.1 Mini acting as a judge to select between pairs of explanations or to reject them if both were equally good or bad. This gives 942 preference/rejection pairs for ORPO training.

## 2 Model Comparison

**Online Evaluation.** Given the subjective nature of user satisfaction, we evaluated the four models described in the above section using a user-based approach. Pre-computed recommendations with explanations were integrated in a lightweight web application[7], enabling pairwise comparison between explanations from different models, similarly to the *Chatbot Arena* platform [6]. We then conducted an evaluation campaign with computer science researchers. Our results ($n = 216$) show no statistically significant difference between the models, although Gemini 2.5 Flash seems slightly better than SFT and ORPO models, which are slightly better than base Qwen3 0.6B. However, the main insight is the poor overall user satisfaction with explanations, regardless of the model, which was confirmed by some participants who were interviewed afterwards.
**Heuristic-based Analysis.** We analyzed the explanations along two dimensions: syntactic and stylistic alignment with the teacher model. For syntactic alignment, measured with ROUGE-L [7], both SFT and ORPO explanations align more closely with the teacher model than the baseline, with ORPO slightly outperforming SFT. For stylistic alignment, lexicon-based analysis reveals that the baseline and the teacher's explanations contained a similar number of domain-specific movie terms, whereas such terms were significantly less frequent in the SFT and ORPO explanations. The baseline explanations retain more residual technical terms from the input explanations than those of ORPO, SFT, and the teacher model. Meanwhile, the teacher model uses more subjective and evaluative language, followed by ORPO, SFT, and the baseline models. Overall, these results suggest that the ORPO explanations model achieves better alignment with the teacher explanations, both syntactically and stylistically.

## 3 Insights & Future Work

The primary objective of this exploratory research was to implement a functional pipeline that enables preliminary experiments. This lays a good foundation for improvement through the integration of related work [8, 9]. The results of the online evaluation are mixed. In particular, they emphasize that designing explanations to improve user satisfaction is not a trivial task. It must be the subject of a dedicated design phase before any RLAIF process. Finally, heuristic analysis has proven insightful for comparing models and enhancing our understanding of the process of aligning fine-tuned models. This reinforces our conviction in the value of combining complementary methods to obtain a comprehensive view.

[7] https://tsw2025-flaskapp-tsw2025-flask-app.apps.factory.trail.ac/

**Acknowledgments.** This research was supported by the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche). We gratefully acknowledge the participants for their valuable contributions to this study.

## References

1. J. Albert, M. Balfroid, M. Doh, J. Bogaert, L. La Fisca, L. De Vos, B. Renard, V. Stragier, and E. Jean, "User preferences for large language model versus template-based explanations of movie recommendations: A pilot study," *arXiv preprint arXiv:2409.06297*, 2024.
2. N. Tintarev and J. Masthoff, "Explaining recommendations: Design and evaluation," in *Recommender systems handbook*, pp. 353–382, Springer, 2015.
3. Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosiute, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional AI: harmlessness from AI feedback," *CoRR*, vol. abs/2212.08073, 2022.
4. F. Yang, N. Liu, S. Wang, and X. Hu, "Towards interpretation of recommender systems with sorted explanation paths," in *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 667–676, IEEE, 2018.
5. J. Hong, N. Lee, and J. Thorne, "Orpo: Monolithic preference optimization without reference model," *arXiv preprint arXiv:2403.07691*, 2024.
6. W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot arena: An open platform for evaluating llms by human preference," 2024.
7. C. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain* (D. Scott, W. Daelemans, and M. A. Walker, eds.), pp. 605–612, ACL, 2004.
8. Y. Luo, M. Cheng, H. Zhang, J. Lu, and E. Chen, "Unlocking the potential of large language models for explainable recommendations," in *International Conference on Database Systems for Advanced Applications*, pp. 286–303, Springer, 2024.
9. M. Yang, M. Zhu, Y. Wang, L. Chen, Y. Zhao, X. Wang, B. Han, X. Zheng, and J. Yin, "Fine-tuning large language model based explainable recommendation with explainable quality reward," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada* (M. J. Wooldridge, J. G. Dy, and S. Natarajan, eds.), pp. 9250–9259, AAAI Press, 2024.