TRUTH: Teaching LLMs to Rerank for Truth in Misinformation Detection

Hao Yu
Mila - Quebec AI Institute
McGill University
hao.yu2@mail.mcgill.ca

Shenyang Huang
University of Oxford
McGill University

Zachary Yang
Mila - Quebec AI Institute
McGill University

Maximilian Puelma Touzel
Mila - Quebec AI Institute
Université de Montréal

Kellin Pelrine
Mila - Quebec AI Institute
McGill University

Jean-François Godbout
Mila - Quebec AI Institute
Université de Montréal

Reihaneh Rabbany
Mila - Quebec AI Institute
McGill University

Abstract

Misinformation detection presents a significant challenge due to its knowledge-intensive and reasoning-intensive nature. While Retrieval-Augmented Generation (RAG) systems offer a promising direction, the effectiveness of their retrieval and reranking components is crucial. This paper introduces TRUTH, a novel reranking approach designed for domain adaptation, specifically for misinformation detection, which employs a two-stage training methodology: Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO). We demonstrate that our 1B parameter TRUTH model achieves strong performance comparable to 7B models on established misinformation benchmarks such as FEVER and Canadian bilingual news datasets, improving retrieval quality and positively impacting downstream task accuracy. Our findings highlight the efficacy of combining SFT for broad knowledge acquisition and domain adaptation with DPO for nuanced reasoning alignment in developing efficient and effective rerankers for complex, knowledge-intensive tasks. Datasets and code will be available with the camera-ready version of the paper.

1 Introduction

Misinformation detection is a knowledge-intensive and reasoning-intensive task (Petroni et al., 2021; SU et al., 2025) that involves knowledge updates and complex reasoning decisions. While high-performance Large Language Models (LLMs) have achieved strong results in translation, natural language understanding, and question answering, they often lack efficient mechanisms to update knowledge needed for verifying recent information. Retrieval-Augmented Generation (RAG) has emerged as a powerful framework for enhancing LLMs by retrieving external knowledge. By incorporating relevant information from a corpus into the generation process, RAG systems significantly improve the factuality and reliability of LLM outputs (Izacard et al., 2023; Zamani & Bendersky, 2024). As shown in Figure 1, a typical retrieval component of a RAG system includes a fusion stage. This stage aggregates and refines retrieved documents from multiple sources to ensure the most relevant information is presented to the subsequent reasoning stage, thereby avoiding cognitive overload in decision-making.

However, current fusion approaches, particularly rerankers, face significant challenges. Firstly, not all retrieved information is beneficial; irrelevant or misleading content can

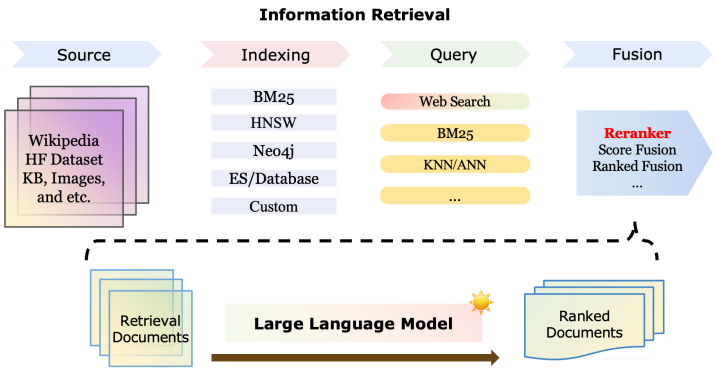

Figure 1: Information retrieval component in a RAG system. The Fusion stage (including rerankers) is critical for gathering and optimizing retrieved documents before generation.

degrade performance, which is especially problematic in misinformation detection where factual accuracy is paramount (Akhtar et al., 2024; Schlichtkrull et al., 2024). Secondly, fact-checking often requires multi-hop reasoning capabilities (Yang et al., 2024), which standard retrieval methods struggle to support. Thirdly, existing encoder-based rerankers often cannot adapt to domain-specific requirements and user preferences (Fernandes et al., 2023). While LLM-based rerankers show promise for reasoning-intensive scenarios, their application to misinformation detection from customized sources remains underexplored.

Our main contributions are as follows:

- We propose TRUTH, a two-stage training process that includes SFT diverse constructed training datasets from general QA benchmarks, and DPO with synthetic reasoning traces.
- We demonstrate the effectiveness of TRUTH through extensive experiments on misinformation detection datasets, including FEVER, Multihop-RAG, and Canada News (EN/FR), showing improvements in both retrieval metrics and downstream task performance.

2 Related Work

2.1 Evaluation of Reranking Models

Our work, TRUTH, builds on advances in RAG systems, reranking techniques, human feedback integration, and misinformation detection strategies. In RAG systems, reranking refines retrieved documents to improve LLM performance (Izacard et al., 2023; Zamani & Bendersky, 2024; Wang et al., 2024). Rerankers have evolved from pointwise (e.g., MonoT5 (Nogueira et al., 2020)) to pairwise (e.g., RankNet (Burges et al., 2005), DuoT5 (Pradeep et al., 2021)), and now to listwise models (e.g., RankT5 (Zhuang et al., 2023), RankGPT (Sun et al., 2023), RankZephyr (Pradeep et al., 2023), FIRST (Reddy et al., 2024)) that assess document sets holistically. We adopt a listwise approach in TRUTH, tailored for misinformation detection via a domain-adaptive two-stage training. While recent LLM-based rerankers (e.g., RankFlow (Jin et al., 2025), ListConRanker) show strong general performance (Ma et al., 2023; Qin et al., 2023; Ren et al., 2024; Chen et al., 2023), they often overlook the challenges of deceptive or subtly false content—gaps TRUTH aims to address.

2.2 Reasoning in Reranking

The direct integration of preference feedback into model training has become increasingly important. Direct Preference Optimization (DPO) (Rafailov et al., 2023), which we employ in our second stage, offers an effective, RL-free method to align models with preferences, outperforming traditional RL-based approaches in several contexts. This aligns with a broader trend of using human or AI-generated feedback to refine NLP models (Li et al., 2016; Hancock et al., 2019; Li et al., 2022; Fernandes et al., 2023; Bai et al., 2024; Xu et al., 2022; Dubois et al., 2023). While some recent works like Re3val (Song et al., 2024a) use reinforcement learning with feedback for generative retrieval, and PRO (Song et al., 2024b) extends pairwise contrasts for preference ranking, our contribution lies in using DPO with

generated *reasoning traces*. This allows TRUTHto learn a ranking based not only on final outcomes but also to internalize aspects of the reasoning process that lead to correct evidence prioritization.

3 Methodology

3.1 Stage 1: Supervised Fine-Tuning

The first stage of the training pipeline focuses on establishing ranking capabilities through SFT on a diverse and carefully curated dataset. This stage is critical for building a robust foundation before introducing reasoning-based preference optimization due to the cold start problem.

Dataset Acquisition We collect data from multiple sources to ensure both coverage and diversity. These datasets include:

- **Base:** The RankZephyr dataset (Pradeep et al., 2023) [1], providing around 40,000 high-quality alphabetic ranking examples.
- **Extended:** Datasets such as MuSiQue (Trivedi et al., 2022), 2WikiMultihopQA (Ho et al., 2020), TriviaQA (Joshi et al., 2017), ChroniclingAmericaQA (Piryani et al., 2024) retrieved with BM25 (Robertson et al., 2009) to introduce task-related and complex scenarios.
- **Domain (Misinformation):** MultiHop-RAG (Tang & Yang, 2024), Canada News (EN/FR), and FEVER (Thorne et al., 2018), targeting fact verification tasks.
  The Canada News (EN/FR) datasets are curated from Canadian government websites using methods detailed in Thibault et al. (2024), with synthetic examples generated with the assistance of GPT-4o-mini[2]. Details on the prompts are provided in Appendix A.2.

Dataset Processing For each query across the datasets, we initially retrieved up to the top 100 related passages using BM25 (Robertson et al., 2009). In some cases, pre-retrieved passages were sourced from the Rankify library (Abdallah et al., 2025). All datasets underwent quality control: we removed duplicate entries and passages that were too short to be informative. We ensured that each training example contained at least one "golden" passage (i.e., a passage containing the answer or critical evidence). During the SFT training phase, to manage context window limitations and improve training efficiency, we sampled 15-20 passages per query. This sampling process guaranteed the inclusion of all identified golden passages for that query.

Data Formatting Each training example for the listwise reranker is formatted with numerical passage identifiers:

```
[1] passage 1 \n[2] passage 2 \n ... [n] passage n
```

This numerical identification scheme was adopted for consistency across all datasets and to support potentially longer context windows with more passages, differing from RankZephyr's original alphabetical identifiers ([A], [B], ...). The complete formatting for SFT and DPO stages is illustrated by examples in Appendix C.1 and is implied by the SFT output format, which is a ranked list of these numerical identifiers.

3.2 Stage 2: Direct Preference Optimization (DPO)

After establishing fundamental ranking capabilities through SFT, we employed DPO to enhance the model's reasoning-based ranking abilities. DPO offers a mathematically principled alignment method that bypasses the need for an explicit reward model, directly optimizing the policy based on preference pairs. More details on DPO are in Appendix B.1.

Reasoning Traces and Preference Pair Construction To train with DPO, we require pairs of preferred (chosen) and dispreferred (rejected) responses as shown in Appendix **?**.

[1] rryisthebest/rank_zephyr_training_data_alpha
[2] GPT-4o-mini

We constructed a reasoning-focused dataset using the misinformation datasets (FEVER, MultiHop-RAG, Canada News EN/FR) from the SFT stage. The process leveraged the 'o4-mini-2025-04-16' model (referred to as o4-mini) and involved the following steps:

1. **Candidate selection**: We selected training queries where BM25 retrieved but misranked.
2. **Reasoning extraction**: Using o4-mini to generate ranking reason without gold evidence.
3. **Reasoning refinement**: We then hint o4-mini with the gold evidence to refine reasoning.
4. **Formatting**: Reasoning and answers were wrapped in `<think>` and `<answer>` tags.
5. **Contrastive pairing**: Pair the formatted and refined responses with empty reasoning tags and randomized passage ranking that deliberately places golden passages lower.

The complete prompt templates used are presented in Appendix C.2. This methodical approach yielded high-quality reasoning examples across all misinformation datasets and verified by researchers using sampled examples. Finally, the DPO dataset comprised 2,491 training and 277 test examples.

4 Results and Discussion

We present a comprehensive analysis of our reranking models, evaluating their performance on reranking quality and downstream misinformation detection tasks across five diverse test datasets: FEVER (dev) (Thorne et al., 2018), FEVER (dev2) (Thorne et al., 2018), MultiHop-RAG (Tang & Yang, 2024), Canada News EN, and Canada News FR. We report Reranking Mean Reciprocal Rank (Rk. MRR@10), downstream task Accuracy (Task Acc.), and F1-score (Task F1). For misinformation detection, to align with the FEVER dataset, the classification labels are "SUPPORT", "REFUTE", and "NOT ENOUGH INFORMATION". BM25 (Robertson et al., 2009) and RankZephyr(7B) (Pradeep et al., 2023) are selected as the comparison baselines. We first retrieve the top 100 passages using BM25, then rerank them with our rerankers. The top 10 passages are selected and passed to LLMs[3] along with the task prompt described in Appendix C.3 to make the final three-class classification.

4.1 Main Performance Comparison

Model	FEVER (dev)			FEVER (dev2)			MultiHop (test)			Canada-News-EN			Canada-News-FR			Average		
	Rk. MRR	Task Acc.	Task F1	Rk. MRR	Task Acc.	Task F1	Rk. MRR	Task Acc.	Task F1	Rk. MRR	Task Acc.	Task F1	Rk. MRR	Task Acc.	Task F1	Rk. MRR	Task Acc.	Task F1
BM25	0.480	65.5	65.8	0.465	55.5	55.5	**0.792**	54.0	62.6	0.803	79.3	79.1	0.725	61.5	62.4	0.653	63.2	65.1
RankZephyr	**0.848**	65.5	64.5	**0.874**	52.0	51.8	0.767	44.5	54.9	0.815	71.0	71.0	0.710	64.0	64.2	**0.803**	59.4	61.3
Llama3.2 1B IT + Base	0.666	70.0	68.4	0.597	57.0	57.1	0.762	35.5	49.6	0.812	86.5	86.8	0.762	66.5	66.9	0.720	63.1	65.8
Llama3.2 1B IT + Base + Misinfor	0.796	73.5	72.1	0.734	56.5	56.3	0.792	38.0	52.4	0.833	86.0	86.3	**0.832**	73.0	73.3	0.798	65.4	68.1
Llama3.2 1B IT + All	0.759	72.5	71.3	0.732	**58.0**	**57.8**	0.751	39.0	53.8	0.833	**88.1**	**88.3**	0.793	**75.5**	**75.8**	0.774	**66.6**	**69.4**
Llama3.2 1B IT + All + DPO	0.767	**74.5**	**73.3**	0.737	57.0	56.8	0.742	**41.5**	**56.4**	**0.835**	82.4	82.8	0.824	68.5	68.8	0.781	64.8	67.6

Table 1: Model performance comparison across datasets. Best scores are **bolded**.

Table 1 shows that SFT consistently improves both reranking and downstream task performance compared to the BM25 baseline. For example, Llama3.2 1B IT + Base improves average Task Accuracy by nearly 2 percentage points over BM25. Adding misinformation-specific data during SFT (Llama3.2 1B IT + Base + Misinfor) further boosts reranking metrics (average Rk. MRR@10 from 0.720 to 0.798) and also task accuracy (average Task Acc. from 63.1% to 65.4%).

Our most comprehensive SFT model, Llama3.2 1B IT + All, which incorporates a wider range of SFT data, achieves the highest average task accuracy (66.6%) and F1-score (69.4%) among all tested models. This highlights the benefit of diverse SFT data.

The addition of DPO on top of the comprehensive SFT model (Llama3.2 1B IT + All + DPO) leads to the competitive average Reranking MRR@10 (0.781) scores among our fine-tuned models. While its average task accuracy is slightly lower than Llama3.2 1B IT + All, it shows strong performance on FEVER (dev), achieving the highest task accuracy (74.5%) on this challenging fact verification dataset. This suggests that DPO is particularly beneficial for

[3] Qwen/Qwen3-32B-AWQ

tasks requiring intensive reasoning and evidence selection, even if the average uplift across all datasets is not always the highest.

Compared to the larger 7B model RankZephyr, our 1B-parameter models Llama3.2 1B IT + All and Llama3.2 1B IT + All + DPO demonstrate competitive performance. Notably, Llama3.2 1B IT + All outperforms RankZephyr in both average task accuracy and F1 score. To ensure a fair comparison by controlling for differences in base models, we trained Llama3.2 1B IT + Base using the same dataset and pipeline as RankZephyr. With the addition of more data and DPO, our enhanced models consistently outperform Llama3.2 1B IT + Base across all downstream tasks.

4.2 Correlation Between Reranking and Task Performance

To understand the relationship between improvements in reranking and downstream task performance, we analyzed the correlation between various metrics.

Metric	Retrieval MRR@10	Reranking MRR@10	Task Accuracy	Task F1
Retrieval MRR@10	1.000	0.540	0.044	0.303
Reranking MRR@10	0.540	1.000	0.214	0.344
Task Accuracy	0.044	0.214	1.000	0.952
Task F1	0.303	0.344	0.952	1.000

Table 2: Correlation matrix between retrieval, reranking, and task performance metrics. Values are Pearson correlation coefficients based on results across all models and datasets. As observed in Table 2, Reranking MRR@10 shows a positive correlation with Task Accuracy (Pearson r = 0.214) and Task F1 (Pearson r = 0.344). This suggests that improvements in the reranker's ability to place relevant documents higher in the list generally translate to better downstream task performance. Notably, the correlation of Task Accuracy with Reranking MRR@10 (0.214) is higher than with Retrieval MRR@10 (0.044), indicating that the reranking stage plays a more crucial role in influencing downstream success than initial retrieval quality alone within the top documents processed by the reranker. This underscores the value of the reranking step. However, the correlations are not extremely strong, implying that other factors, such as the complexity of the reasoning required by the task and the capabilities of the generation model, also play significant roles.

4.3 Impact of Reranking at Different Top-K Levels

Next, we further investigate how "reranking" impacts accuracy at various cut-off points (Top-K hit accuracy). Figure 2 illustrates the trend of accuracy (defined as the presence of at least one relevant document) for different values of K, comparing BM25 with our best-performing fine-tuned models.

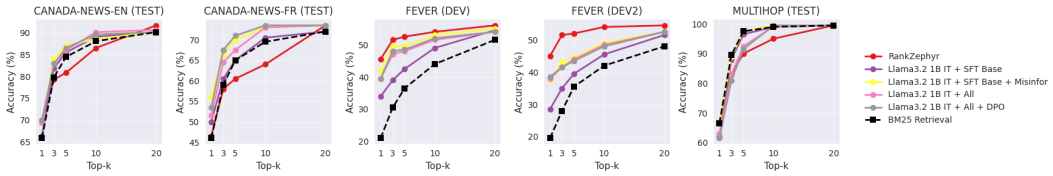

Figure 2: Accuracy@K for BM25, the baseline model, and fine-tuned models. This plot shows the percentage of queries for which at least one relevant document is found within the top K passages across all datasets.

Figure 2 clearly shows that the most significant improvements from reranking are observed at smaller values of K (e.g., K=1, 3, 5, 10). For instance, the gap between BM25 and the fine-tuned models like Llama3.2 1B IT + All or Llama3.2 1B IT + All + DPO is largest for K=1 and K=3. This is critical because downstream LLMs often only process a small number of passages due to context window limitations or efficiency considerations. As K increases, the accuracy of BM25 naturally improves, and the incremental benefit of reranking diminishes. This pattern is expected: rerankers excel at precisely ordering the very top documents.

While not shown directly in this plot, the downstream task accuracy (which typically uses Top-5 or Top-10 passages) also benefits from this early precision.

5 Conclusion

We introduced TRUTH, a two-stage training approach (SFT followed by DPO with AI-generated reasoning) to enhance 1B-parameter rerankers for misinformation detection. Experiments show that SFT establishes strong ranking baselines, while DPO further refines reasoning, particularly for complex queries. Our fine-tuned 1B models achieve competitive performance (e.g., "Llama 3.2 1B IT + SFT All" with a 66.6% average task accuracy) against larger models, demonstrating the efficacy of TRUTH for developing efficient, reasoning-aware rerankers. For future work, we plan to explore more advanced methods for generating diverse and high-fidelity reasoning traces for DPO training, potentially incorporating self-critique mechanisms. Additionally, investigating techniques to further improve cross-lingual transfer and adaptability to new, unseen domains remains a key direction.

Acknowledgments

We gratefully acknowledge the support of IVADO from the Canada First Research Excellence Fund (CFREF) for developing robust, reasoning, and responsible artificial intelligence (IAR^3), specifically under Regroupement 4: Implementation and Governance, Project: Quality Data Governance ("Fonds d'excellence en recherche Apogée Canada pour développer une intelligence artificielle robuste, raisonnante, et responsable (IAR^3). Regroupement 4, Mise en oeuvre et gouvernance. Projet Gouvernance des données de qualité"). Additionally, we are grateful for the computational resources provided by the Mila cluster and the Compute Canada Tamia cluster, which were essential for conducting this research.

References

Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation. February 2025. doi: 10.48550/ARXIV.2502.02464.

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. Ev2r: Evaluating evidence retrieval in automated fact-checking. November 2024. doi: 10.48550/ARXIV.2411.05375.

Yu Bai, Yukai Miao, Li Chen, Dawei Wang, Dan Li, Yanyu Ren, Hongtao Xie, Ce Yang, and Xuhui Cai. Pistis-rag: Enhancing retrieval-augmented generation with human feedback. June 2024. doi: 10.48550/ARXIV.2407.00072.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Ziyi Chen, Jize Jiang, Daqian Zuo, Heyi Tao, Jun Yang, and Yuxiang Wei. Efficient title reranker for fast and improved knowledge-intense nlp. December 2023. doi: 10.48550/ARXIV.2312.12430.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 30039–30069. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5fc47800ee5b30b8777fdd30abcaaf3b-Paper-Conference.pdf.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:

1643–1668, 12 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00626. URL https://doi.org/10.1162/tacl_a_00626.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*, 2019.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023. URL http://jmlr.org/papers/v24/23-0037.html.

Can Jin, Hongwu Peng, Anxiang Zhang, Nuo Chen, Jiahui Zhao, Xi Xie, Kuangzheng Li, Shuya Feng, Kai Zhong, Caiwen Ding, and Dimitris N. Metaxas. Rankflow: A multi-role collaborative reranking workflow utilizing large language models. February 2025. doi: 10.48550/ARXIV.2502.00709.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*, 2016.

Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie C. K. Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. *Findings of the Association for Computational Linguistics: ACL (2022) 926-937*, pp. 926–937, April 2022. doi: 10.18653/v1/2022.findings-acl.75.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. Zero-shot listwise document reranking with a large language model. May 2023. doi: 10.48550/ARXIV.2305.02156.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. March 2020. doi: 10.48550/ARXIV.2003.06713.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 2038–2048, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657891. URL https://doi.org/10.1145/3626772.3657891.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. January 2021. doi: 10.48550/ARXIV.2101.05667.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! December 2023. doi: 10.48550/ARXIV.2312.02724.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. June 2023. doi: 10.48550/ARXIV.2306.17563.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Technical report, December 2023. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs] type: article.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. First: Faster improved listwise reranking with single token decoding. June 2024. doi: 10.48550/ARXIV.2406.15657.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. Self-calibrated listwise reranking with large language models. November 2024. doi: 10.48550/ARXIV.2411.04602.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36, 2024.

EuiYul Song, Sangryul Kim, Haeju Lee, Joonkee Kim, and James Thorne. Re3val: Reinforced and reranked generative retrieval. January 2024a. doi: 10.48550/ARXIV.2401.16979.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18990–18998, Mar. 2024b. doi: 10.1609/aaai.v38i17.29865. URL https://ojs.aaai.org/index.php/AAAI/article/view/29865.

Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ykuc5q381b.

Weiwei Sun, Zheng Chen, Xinyu Ma, Lingyong Yan, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Instruction distillation makes large language models efficient zero-shot rankers. November 2023. doi: 10.48550/ARXIV.2311.01555.

Yixuan Tang and Yi Yang. Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t4eB3zYWBK.

Camille Thibault, Jacob-Junqi Tian, Gabrielle Peloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. A guide to misinformation detection data and evaluation. November 2024. doi: 10.48550/ARXIV.2411.05060.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Zihan Wang, Xuri Ge, Joemon M. Jose, Haitao Yu, Weizhi Ma, Zhaochun Ren, and Xin Xin. R3ag: First workshop on refined and reliable retrieval augmented generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, pp. 307–310, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707247. doi: 10.1145/3673791.3698435. URL https://doi.org/10.1145/3673791.3698435.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. August 2022. doi: 10.48550/ARXIV.2208.03270.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? November 2024. doi: 10.48550/ARXIV.2411.16679.

Hamed Zamani and Michael Bendersky. Stochastic rag: End-to-end retrieval-augmented generation through expected utility maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 2641–2646, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657923. URL https://doi.org/10.1145/3626772.3657923.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2308–2313, 2023.

A Dataset

A.1 Training Dataset Analysis

Category	Dataset	Retrieval Model	Original Count	Final Count / Selected
Origin	RankZephyr	-	39,912	39,912
Extended	musique (dev)	BM25	2,417	998 / 0
	musique (train)	BM25	2,417	8137 / 8000
	2WikiMultihopQA (train)	BM25	14,999	8,655 / 8000
	2WikiMultihopQA (dev)	BM25	12,576	7,693 / 0
	TriviaQA (dev)	BM25	8,837	7,387 / 0
	TriviaQA (train)	ColBERT	78,785	65,898 / 8000
	ChroniclingAmericaQA (val)	BM25	24,111	7,994 / 7994
	ArchivialQA (val)	BM25	-	28,373 / 8000
	HotpotQA (val)	Contriever	-	3,518 / 0
	HotpotQA (val)	DPR	-	4,138 / 4,138
	Subtotal		*141,725*	*/ 40,072*
Misinfor	MultiHop (train)	BM25/BGE	940	938
	Canada News EN (train)	BM25	896	866
	Canada News FR (train)	BM25	1,140	908
	FEVER (train)	BM25	300	182
	Subtotal		*3,276*	*2,894*
Train Total			163,262	**67,023**
	MultiHop (test)	BM25/BGE	202	200
	Canada News EN (train)	BM25	193	193
	Canada News FR (train)	BM25	245	200
	FEVER (dev)	BM25	9,999	200
	FEVER 2.0 Dev (dev2)	BM25	1,174	200
Test Total			11,813	**1997**

Table 3: Detailed dataset composition for Supervised Fine-Tuning and evaluation. The final count represents the number of examples after filtering for quality and relevance.

A.2 Canada News Dataset

Data Source Links

- https://api.io.canada.ca/io-server/gc/news/en/v2 [EN]
- https://www.ourcommons.ca/en/rss-feed [EN]
- https://sencanada.ca/umbraco/surface/NewsAjax/GetNews [EN/FR]
- https://www.quebec.ca/en/news/search [EN]
- https://www.quebec.ca/nouvelles/rechercher [FR]
- https://api.news.ontario.ca/api/v1/releases [EN/FR]

Canada News Dataset Generation Prompt The prompt used for generating True, False, and Unanswerable claims for the Canada News datasets:

```
A "claim" is a statement or assertion made within a text expressing a belief, opinion, or
↪  fact. Given evidence from the original context, please extract one claim and create
↪  related variants.

Here are some example claims from other datasets to give you an idea of what we're
↪  looking for:
{claim_example}

For the provided evidence, please generate THREE different types of claims:

1. A TRUE CLAIM that accurately represents information in the evidence
2. A FALSE CLAIM that appears credible but contains misinformation by using one of these
↪  techniques:
   - Exaggerate or overstate consequences
   - Invert meaning or state the opposite
   - Change important details (dates, numbers, locations, people)
   - Add false connections or motives
   - Remove critical context
   - Falsely attribute information
   - Create claims that have insufficient verification details

3. An UNANSWERABLE CLAIM that:
   - Is related to the evidence topic
   - Asks for specific information that is not provided in the evidence
   - Cannot be verified using only the given evidence
   - Appears reasonable but requires additional information to verify

Please format your response as follows:
##Evidence: {original context}
##True Claim: {accurate claim based on evidence}
##True Claim Target: {entity the true claim is about}
##True Claim Topic: {topic of the true claim}
##False Claim: {deliberately inaccurate claim}
##False Claim Target: {entity the false claim is about}
##False Claim Topic: {topic of the false claim}
##Unanswerable Claim: {claim that cannot be verified from evidence}
##Unanswerable Claim Target: {entity the unanswerable claim is about}
##Unanswerable Claim Topic: {topic of the unanswerable claim}
##Why Unanswerable: {A brief explanation of why this claim cannot be answered from the
↪  evidence}

Example:
##Evidence: The government announced a $5 million investment in renewable energy
↪  projects across the country.
##True Claim: The government is investing $5 million in renewable energy projects
↪  nationwide.
##True Claim Target: government
##True Claim Topic: renewable energy investment
##False Claim: The government secretly diverted $5 million meant for healthcare to fund
↪  experimental renewable energy projects that were later abandoned.
```

```
##False Claim Target: government
##False Claim Topic: fund diversion
##Unanswerable Claim: The government's $5 million investment in renewable energy will
↪  create over 10,000 new jobs nationwide.
##Unanswerable Claim Target: government investment
##Unanswerable Claim Topic: job creation
##Why Unanswerable: The evidence doesn't mention any job creation figures or employment
↪  impact from the investment.

Now, please analyze the following text:
##Input: {original context}

# Pydantic model for combined claims (Canada News)
class CombinedClaims(BaseModel):
    evidence: str
    true_claim: str
    true_claim_target: str
    true_claim_topic: str
    false_claim: str
    false_claim_target: str
    false_claim_topic: str
    unanswerable_claim: str
    unanswerable_claim_target: str
    unanswerable_claim_topic: str
    why_unanswerable: str
```

B Additional Details on Direct Preference Optimization

B.1 Direct Preference Optimization Introduction

Direct Preference Optimization (DPO) (Rafailov et al., 2023) has emerged as an effective RL-free technique for aligning models with human preferences. Instead of explicitly training a reward model and then using RL, DPO leverages a mapping between reward functions and optimal policies. It directly optimizes the language model policy using a simple binary cross-entropy loss on preference pairs (x, y_w, y_l), where y_w is the preferred and y_l is the dispreferred completion for prompt x. The DPO loss is defined as:

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

where π_θ is the policy being optimized, π_{ref} is a reference policy (usually the SFT model), β controls the deviation from the reference policy, and σ is the logistic function. This approach implicitly optimizes a reward function while being computationally lightweight and stable.

C Data Examples and Prompt Templates

C.1 Example of SFT and DPO Training Pair

This example shows the structure of preference pairs, where the chosen response includes correct reasoning and ranking, while the rejected response lacks reasoning and ranks incorrectly.

C.2 DPO Dataset Construction Prompts

The following prompts were used in the DPO training pipeline.

Initial Thinking Prompt Used to obtain initial reasoning without gold evidence.

```
System: You are RankLLM, an intelligent assistant that can rank passages based on their
↪  relevance to the query.
```

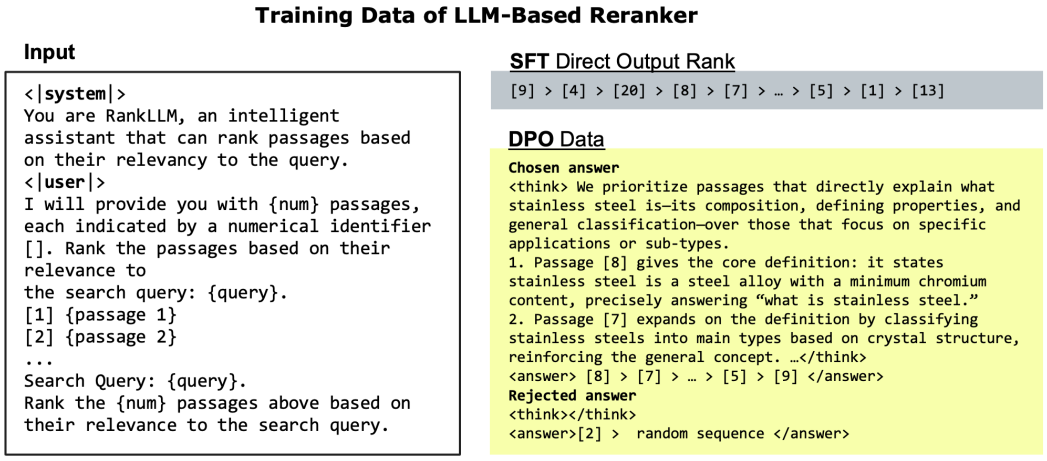

Figure 3: Example of a training data instance for SFT and a preference pair for DPO.

```
User: I will provide you with {num_contexts} passages, each indicated by a numerical
↪  identifier [].
Rank the passages based on their relevance to the search query: {query}

{contexts}

Search Query: {query}

Think carefully about the relevance of each passage to the query.
Explain your reasoning process in detail, and then provide your final ranking.
For the final ranking, list all passages in descending order of relevance using the
↪  format [N] > [M] > etc.
```

In-Language Thinking Refinement Prompt Used to refine reasoning in the query's original language, this was the primary method for generating "chosen" responses for DPO.

```
System: You are RankLLM, an intelligent assistant that can rank passages based on their
↪  relevance to the query.

User: I received the following thinking process and ranking for this search query:
↪  {query}

Initial thinking and ranking:
{initial_thinking_response}

The passages that actually contain the answer are: {golden_ids_str}

Please refine the thinking process to focus on why these passages are most relevant to
↪  the query.
Format your thinking in the same language as the query ({language}).

Format your response with the thinking part wrapped in <think></think> tags and the final
↪  ranking wrapped in <answer></answer> tags.

The final ranking should be in the same language as the query.

The final ranking should include all passages in descending order of relevance using the
↪  format [N] > [M] > etc.
```

C.3 Misinformation Detection Prompt

The following prompt is used for the final claim verification task, instructing the LLM on how to classify claims based on the provided evidence.

```
# Claim Verification Task

Analyze the provided evidence to determine whether it supports, refutes, or is
↪  insufficient to evaluate the given claim.

## Classification Rules

**SUPPORT**: All evidence consistently supports the claim with sufficient detail
**REFUTE**: Evidence contradicts or conflicts with the claim
**UNANSWERABLE**: Evidence is insufficient, unclear, or contains conflicting information

## Process

1. **Examine Evidence**: Review each piece of evidence for relevance and accuracy
2. **Compare to Claim**: Check if evidence aligns with, contradicts, or is unrelated to
↪  the claim
3. **Make Judgment**: Select the appropriate classification based on overall evidence
↪  quality and consistency

## Output Format

**Claim**: [Insert claim here]

**Analysis**: [Brief explanation of how evidence relates to the claim]

**Judgment**: SUPPORT/REFUTE/UNANSWERABLE

---

**Claim**:
{claim}

**Evidence List**:
{evidence}

**Analysis**:

**Judgment**:
```