Training of LLM-Based List-Wise Multilingual Reranker

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

David Ifeoluwa Adelani
Mila - Quebec AI Institute
McGill University & Canada CIFAR AI Chair
david.adelani@mila.quebec

Abstract

Multilingual retrieval-augmented generation (MRAG) systems heavily rely on robust Information Retrieval (IR). Reranking as a key component optimizes the initially retrieved document set to present the most pertinent information to the generative model, addressing context limitations and minimizing hallucinations. We propose an approach that trains Large Language Models (LLMs) as multilingual listwise rerankers through supervised fine-tuning (SFT) on a diverse mixture of multilingual and extended English ranking examples, and enhancing reasoning capabilities through Direct Preference Optimization (DPO) from translated task-specific reasoning processes. Experiments demonstrate that the approach improves accuracy@5 by 20-30% across all six high- medium- and low-resource languages compared to the BM25. The posted training 1B models achieve comparable performance to 7B baseline models while enabling faster inference. Finally, we investigate the effectiveness of different reasoning strategies in DPO with crosslingual and monolingual thinking processes.

1 Introduction

Large Language Models (LLMs) often struggle with factuality, particularly in multilingual contexts with limited training data. RAG systems address this by combining LLMs with external knowledge retrieval, enhancing language performance. In these systems, the IR component is essential, with reranking playing a critical role in refining retrieved documents before decision making ("Fusion" Stage in Figure 1).

This paper focuses on the rerankers in the multilingual setting, a key component that optimizes retrieved content across diverse languages, ensuring the most relevant information is provided to LLMs while maintaining efficiency and performance even with limited computational resources.

Recent advances in reranking have leveraged transformer-based architectures, with LLM-based listwise rerankers showing particular promise for reasoning-intensive scenarios. Despite these advances, multilingual reranking is underexplored and remains challenging.

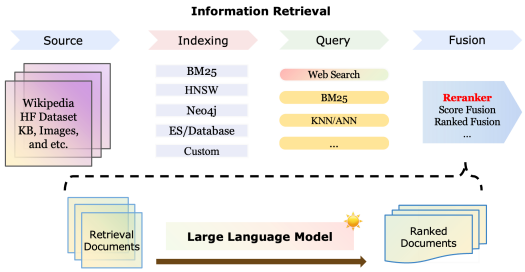

Figure 1: Common stages in information retrieval processes. The last "Fusion" stage is critical for gathering and optimizing retrieved documents before generation.

Our contributions are as follows:

- Construct the training dataset with various English QA datasets with retrieval golden labels and multilingual retrieval datasets with thinking traces from o4-mini[1].
- Firstly propose a two-stage training methodology combining SFT and DPO to enhance the capabilities of the ranking procedure and enable reasoning ability separately.
- Investigate the impact of reasoning strategies of language choice, comparing translated English versus in-language thinking.

2 Related Works

2.1 Multilingual Information Retrieval

Multilingual Information Retrieval (MLIR) extends reranking to cross-language and multiple languages scenarios, presenting unique challenges beyond monolingual retrieval. A key difficulty is producing comparable relevance scores across languages while avoiding language bias – the ten-

[1] OpenAI o4-mini

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 652–663
November 8-9, 2025

dency for retrieval quality to vary by language. (Yang et al., 2024b) found that BM25 rankings for semantically identical queries in different languages diverge significantly, whereas neural models show more consistent behaviours. The other three primary strategies shown in Appendix A.1 have emerged for MLIR reranking also, such as translation pipelines, multilingual pre-trained models and loss-based alignment.

2.2 Reranking with Human Feedback

Integrating human feedback in LLMs has become increasingly important for model alignment with human preferences. The most common approach involves supervised fine-tuning (SFT), where models learn from labelled examples of optimal rankings – highlighted grey in the below Figure 2. (Pradeep et al., 2023)

However, research indicates that simple SFT is insufficient to fully address the challenges presented by complex benchmarks like MIRACL. To overcome these limitations, researchers have incorporated explicit reasoning steps and error feedback during training. Two notable approaches in this direction are: DPO (Rafailov et al., 2023) provides a straightforward method for preference alignment without requiring explicit reward modeling and GRPO (Shao et al., 2024) demonstrated effectively in DeepSeek Math (Shao et al., 2024), which leverages group-wise rewards to improve model performance. Other specialized approaches include Re3val (Song et al., 2024), a reinforced reranking method for generative retrieval, and Preference Ranking Optimization (PRO), which extends DPO to handle preference rankings of arbitrary length. Farinhas et al. 2024 introduced a communication-theoretic perspective, optimizing for information preservation.

2.3 Datasets of MLIR

Evaluation datasets have expanded significantly in recent years. MIRACL (Zhang et al., 2023) provides ad-hoc retrieval queries and relevance judgments in 18 typologically diverse languages using Wikipedia passages. Multi-EuP (Yang et al., 2023) offers European Parliament documents in 24 EU languages with fully parallel queries. BordIRlines (Li et al., 2024) contains queries about disputed territories with aligned passages in 49 languages. For RAG evaluation, NoMIRACL (Thakur et al., 2024) provides human-labelled non-relevant and relevant passage sets to test retrieval robustness across 18 languages. Mr.TyDi (Zhang et al., 2021) is a diverse multilingual benchmark covering eleven typologically distinct languages, designed for monolingual retrieval evaluation. It provides queries, relevance judgments, and training data with negative examples from the top-30 BM25 results.

Input

```
<|system|>
You are RankLLM, an assistant …
<|user|>
[1] {passage 1}\n[2] {passage 2}...
Search Query: {query}.
Rank the {num} passages above based on
their relevance to the search query.
```

SFT Direct Output Rank

```
[9] > [4] > [20] > [8] > [7] > … > [1] > [13]
```

DPO Thinking Preference Pair

```
Chosen answer
<think>
1. Passage [8] gives the core definition: it
states stainless steel is a steel alloy with a
minimum chromium content.
2. Passage [7] expands on the definition by
classifying stainless steels into main types based
…</think>
<answer> [9] > [4] > … > [1] > [13]</answer>
Rejected answer
<think></think>
<answer>[2] > random sequence </answer>
```

Figure 2: Training data example of SFT and DPO.

3 Methodology

This section will introduce our two-stage training pipeline for developing efficient multilingual rerankers. First, we establish foundational ranking capabilities through SFT on a diverse and curated dataset. Then, we enhance reasoning-based ranking capabilities using DPO with structured thinking processes.

3.1 Stage 1: Supervised Fine-Tuning

The first stage of the training pipeline focuses on establishing strong multilingual ranking capabilities through SFT on a diverse and curated dataset.

3.1.1 Dataset Construction and Preparation

We aggregate data from multiple sources to ensure both coverage and diversity. The dataset includes:

- **Base:** The RankZephyr dataset (Pradeep et al., 2023) [2], providing around 40,000 high-quality English ranking examples.
- **English Extended:** Datasets such as MuSiQue (Trivedi et al., 2022), 2WikiMultihopQA (Ho et al., 2020), TriviaQA (Joshi et al., 2017), ChroniclingAmericaQA (Piryani et al., 2024), MultiHop-RAG (Tang and Yang,

[2] https://huggingface.co/datasets/rryisthebest/rank_zephyr_training_data_alpha

2024), Canada News (EN/FR), and FEVER (Thorne et al., 2018) retrieved with BM25 (Robertson et al., 2009) or ColBERT (Khattab and Zaharia, 2020), to introduce task related and complex reasoning scenarios.

- **Multilingual (TyDi (Zhang et al., 2021)):** Arabic, English, Japanese, and Swahili subsets, enabling cross-lingual ranking ability.

All datasets are filtered for quality: we remove duplicates, passages that are too short, and ensure each example contains at least one passage with golden evidence. For TyDi, we sampled 15-20 passages per query, always including golden evidence. Overall, the Table 3 in Appendix summarizes the original and final counts for each dataset after filtering, as well as the retrieval model used. Subtotals are provided for each group.

3.2 Stage 2: Direct Preference Optimization

After establishing fundamental ranking capabilities through SFT, we employed DPO to enhance the models' reasoning-based ranking abilities. DPO offers a mathematically principled alignment approach that bypasses the need for an explicit reward model. Additional technical details about DPO are provided in Appendix A.2.

Reasoning Dataset Construction To develop an effective DPO training corpus for multilingual reasoning, we leveraged o4-mini to construct the first reasoning-focused dataset specifically designed for list-wise ranking across multiple languages. The construction process followed these key steps:

1. **Strategic candidate selection**: We use queries from the TyDi training split where BM25 retrieval successfully included golden evidence passages but failed to rank them.
2. **Reasoning extraction**: We prompted o4-mini to generate detailed reasoning traces for these selected queries without revealing golden evidence information.
3. **Reasoning refinement**: In a second pass, we provided both the initial reasoning and golden evidence information to o4-mini, guiding it to produce improved reasoning that correctly identified the most relevant passages.
4. **Structural formatting**: All content was consistently formatted with reasoning processes enclosed in `<think>...</think>` tags and final rankings in `<answer>...</answer>` tags, creating clear separation between reasoning process and ranking output.

The complete prompt templates used for this reasoning generation are documented in Appendix C. This methodical approach yielded high-quality reasoning examples across all target languages.

Translating Thinking We further investigated two distinct cross-lingual reasoning strategies, as outlined in the following Table 1. The final DPO training corpus follows the preference pair construction example in Figure 2 and comprises 3,267 training and 363 test examples for in-language reasoning, alongside 3,199 training and 359 test examples for translated reasoning.

Strategy	Description
Translated	Request model translates passages into English, conducts reasoning in English, and then ranks.
In-Language	The model maintains the source language throughout both the reasoning and ranking processes.

Table 1: Cross-lingual reasoning strategies used for DPO, prompts are displayed in Appendix C.

4 Experiments

Evaluation Dataset We evaluate reranker models using MIRACL (Zhang et al., 2023), a multilingual information retrieval dataset with queries and relevant passages across 18 languages, focusing on the 6 languages described in Table 3.

Evaluation Metrics We measure performance using Top-k accuracy, noted as $acc@k$, which determines whether at least one relevant document appears in the first k retrieved documents. Report results for $k \in \{1, 3, 5, 10, 20\}$.

Baseline Models

- **BM25**[3]: Standard retrieval model without reranking. For each query, retrieved top 100.
- **RankZephyr** (Pradeep et al., 2023): Listwise reranker based on Zephyr 7B architecture
- **Llama-3.2-1B-Instruct** (Grattafiori et al., 2024)/ **Gemma-3-1b-it** (Gemma Team et al., 2025) SFT: 1B parameter models trained on the same dataset as RankZephyr (Pradeep et al., 2023).

5 Results

5.1 Supervised Fine-Tuning Results

Table 2 presents $acc@5$ across languages, revealing a striking divergence in how architectures respond to multilingual TyDi data. Gemma-3-1B experiences catastrophic performance degradation when

[3]bm25s.github.io

Model	English	French	Arabic	Japanese	Swahili	Yoruba	Non-En Avg
BM25 (No reranking)	62.5	26.0	59.0	54.5	55.0	52.1	49.3
RankZephyr (7B)	**80.5**	**51.5**	74.0	52.5	64.5	63.9	61.3
Gemma-3-1B Origin	63.0	28.0	60.0	58.0	56.5	51.3	50.8
Gemma-3-1B Origin + Extended	60.0	37.5	59.5	44.5	45.5	42.9	46.0
Gemma-3-1B Origin + TyDi	40.0	25.5	40.0	24.5	27.5	21.8	27.9
Gemma-3-1B + All	60.0	42.0	65.0	51.5	51.0	48.7	51.6
Llama-3.2-1B Origin	70.0	37.0	65.0	58.0	59.5	54.6	54.8
Llama-3.2-1B Origin + Extended	74.5	49.5	74.5	61.5	68.5	60.5	62.9
Llama-3.2-1B Origin + TyDi	68.5	41.5	68.5	57.0	64.5	55.5	57.4
Llama-3.2-1B + All	76.0	48.5	74.5	63.5	**69.0**	63.0	63.7
Pure DPO (Translated)	76.5	49.0	74.5	**64.5**	**69.0**	64.7	**64.3**
Pure DPO (In-language)	61.0	26.5	59.0	54.5	55.5	52.1	49.5
Llama-3.2-1B + All + DPO (Translated)	76.5	49.0	74.5	64.0	**69.0**	64.7	64.2
Llama-3.2-1B + All + DPO (In-language)	77.0	49.0	**75.0**	63.5	**69.0**	62.2	63.7

Table 2: Model performance comparison across languages (Acc@5)

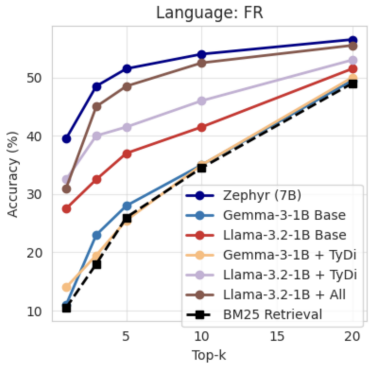

Figure 3: Performance across different top-k values in French. Figure 5 in appendix covers other languages.

trained with TyDi data, with drops of 20-33 points across all languages. In contrast, Llama-3.2-1B shows resilience with the same data, ranging from minimal decline in English (-1.5 points) to gains in Arabic (+3.5 points) and Swahili (+5 points).

Despite similar parameter counts, Llama-3.2-1B consistently outperforms Gemma-3-1B across all languages, with the gap widening when including TyDi data. The best-performing Llama-3.2-1B model approaches or exceeds the much larger RankZephyr (7B) model, delivering improvements over BM25 ranging from 9 to 22.5 points. Performance varies by language, with English and Arabic showing highest accuracy, while Japanese and French present greater challenges. The gain is more pronounced for languages covered in training data (Arabic, Swahili, Japanese) compared to other non-covered languages.

Analyzing retrieval patterns across different k values (Figure 3), improvements are most pronounced at lower k values. The improvement curves flatten as k increases, with most dramatic gains occurring between k=1 and k=5. Japanese and French show more gradual improvement as k increases compared to English and Arabic, suggesting different document relevance distributions.

Moreover, Llama-3.2-1B+All outperforms the larger RankZephyr (7B) model across most lower-resource languages (Arabic, Japanese, Swahili, Yoruba), while RankZephyr maintains an edge in high-resource languages (English, French). This suggests our approach of mixing diverse training data is particularly effective for lower-resource languages, even with smaller models.

5.2 Direct Preference Optimization Results

DPO experiments results from Table 2 reveal clear patterns regarding reasoning strategy and training methodology. Reasoning strategy dramatically affects pure DPO performance. Models trained with in-language thinking regress to baseline BM25 levels across all non-English languages. Conversely, translated thinking (reasoning in English) yields strong improvements comparable to SFT models, suggesting stronger reasoning capabilities in English benefit multilingual reranking.

Combined SFT+DPO approach mitigates reasoning strategy sensitivity. When applied after SFT, both reasoning approaches yield similar results, with in-language thinking showing only slight degradation. The SFT phase provides a foundation that DPO can effectively refine.

6 Conclusion

Our results demonstrate that compact 1B-parameter models can effectively perform multilingual reranking when appropriately trained, with Llama-3.2-1B consistently outperforming Gemma-3-1B, particularly with diverse training data. The dramatic differences between model families in their ability to incorporate multilingual data highlight the importance of architecture in cross-lingual transfer. For deployment scenarios requiring efficiency across multiple languages, carefully trained 1B models offer an attractive alternative to larger 7B models with comparable performance but faster inference.

7 Acknowledgment

We would like to express our gratitude to the researchers whose work laid the foundation for this study. We are particularly thankful for access to computational resources provided by the Mila cluster and Compute Canada GPU infrastructure. David Adelani acknowledges the funding of IVADO and the Canada First Research Excellence Fund.

We also extend our appreciation to McGill University for offering the Multilingual Representation Learning course, which inspired and guided this research. This project originated as the final project for that course and benefited greatly from the knowledge and frameworks presented throughout the semester.

8 Limitations

Despite promising results, our approach faces several important limitations:

Language Coverage While we demonstrate improved performance across six languages, our training focuses primarily on four languages (Arabic, English, Japanese, and Swahili). The generalization to low-resource languages remains challenging, as evidenced by the relatively lower performance gains in Yoruba and French. Future work should incorporate a broader language spectrum during training to better address linguistic diversity.

Reasoning Quality While our DPO approach improves reasoning capabilities, the quality of reasoning varies significantly between languages. The stark difference between translated and in-language reasoning performance suggests that reasoning abilities in non-English languages remain underdeveloped in these models, creating potential fairness issues in deployment scenarios.

GRPO Implementation Challenges Our attempts to implement Group Relative Policy Optimization (GRPO) with language-specific reward functions did not yield stable results, often producing random strings instead of coherent rankings. This suggests fundamental challenges in designing effective reward functions for multilingual reranking tasks, particularly for maintaining language consistency during reasoning. The language-alignment reward function showed promise in concept but requires further research to stabilize training dynamics.

Computational Resources Although our 1B parameter models offer efficiency advantages over larger models, the two-stage training pipeline still requires substantial computational resources, particularly during the DPO phase. This may limit accessibility for research groups with limited infrastructure.

Evaluation Metrics Our evaluation primarily focuses on accuracy@k metrics, which may not fully capture nuanced aspects of ranking quality such as diversity, fairness across demographic groups, or robustness to adversarial queries. The rank-based metrics could be adopted, such as MRR (Mean Reciprocal Rank), MAP@k (Mean Average Precision).

Future work should address these limitations by expanding language coverage, developing more stable GRPO implementations with carefully designed reward functions, and exploring alternative evaluation frameworks that better capture real-world performance considerations across diverse linguistic contexts.

References

Mofetoluwa Adeyemi, Akintunde Oladipo, Ronak Pradeep, and Jimmy Lin. 2024. Zero-shot cross-lingual reranking with large language models for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–656, Bangkok, Thailand. Association for Computational Linguistics.

António Farinhas, Haau-Sing Li, and André F. T. Martins. 2024. Reranking laws for language generation: A communication-theoretic perspective. In *Advances in Neural Information Processing Systems*, volume 37, pages 111074–111105. Curran Associates, Inc.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran

Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Gold-

schlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe,

and Chris Callison-Burch. 2024. BordIRlines: A dataset for evaluating cross-lingual retrieval augmented generation. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 1–13, Miami, Florida, USA. Association for Computational Linguistics.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. 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, page 2038–2048, New York, NY, USA. Association for Computing Machinery.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze!

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

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18990–18998.

Yixuan Tang and Yi Yang. 2024. Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*.

Nandan Thakur, Luiz Bonifacio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. "knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12508–12526, Miami, Florida, USA. Association for Computational Linguistics.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Eugene Yang, Dawn Lawrie, and James Mayfield. 2024a. Distillation for multilingual information retrieval. pages 2368–2373.

Jinrui Yang, Timothy Baldwin, and Trevor Cohn. 2023. Multi-EuP: The multilingual European parliament dataset for analysis of bias in information retrieval. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 282–291, Singapore. Association for Computational Linguistics.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2024b. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 280–292, Miami, Florida, USA. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A Extended Relative Work

A.1 MLIR Reranking pipeline

Translation pipelines convert either queries or documents into a pivot language (typically English) to leverage monolingual rankers. (Adeyemi et al., 2024) evaluated LLM rerankers by translating between English and four African languages, finding that LLMs perform best when operating in English, but cross-lingual setups can approach monolingual effectiveness with sufficiently multilingual models.

Multilingual pre-trained models like mBERT, XLM-R, and multilingual T5 enable direct cross-lingual encoding. Recent work by (Zhang et al., 2024) developed mGTE, a new long-context (8192

tokens) multilingual encoder with a contrastively trained reranker that achieves SOTA performance across multiple languages.

Contrastive and loss-based alignment techniques explicitly align language representations. (Yang et al., 2024a) proposed Multilingual Translate-Distill (MTD), which trains a multilingual dual encoder using translation and teacher-student distillation to ensure consistently scored documents across languages.

A.2 Direct Preference Optimization

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.

B Experiments Details and Results

B.1 Data Statistic

Category	Dataset	Retrieval Model	Original Count	Final Count
Origin	RankZephyr	-	39,912	39,912
Extended	musique (dev)	BM25	2,417	998
	2WikiMultihopQA (train)	BM25	14,999	8,655
	2WikiMultihopQA (dev)	BM25	12,576	7,693
	TriviaQA (dev)	BM25	8,837	7,387
	TriviaQA (train)	ColBERT	78,785	67,711
	ChroniclingAmericaQA (val)	BM25	24,111	7,994
	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		*144,701*	*103,332*
Multilingual (TyDi)	Arabic	BM25	12,335	7,484
	English	BM25	3,547	3,119
	Japanese	BM25	3,697	3,364
	Swahili	BM25	2,072	1,888
	Subtotal		*21,651*	*15,855*
Train Total			212,051	**160,206**
Multilingual Evaluation (MIRACL)	Arabic (ar)	BM25	2,896	200
	English (en)	BM25	799	200
	Japanese (ja)	BM25	860	200
	Swahili (sw)	BM25	482	200
	Yoruba (yo)	BM25	119	119
	French (fr)	BM25	343	200
	Test Total		5,499	**1,119**

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.

B.2 Finetuning Setup

For training Llama-3.2-1B-SFT and Gemma-3-1B-it SFT, we follow RankZephyr (Pradeep et al., 2023) with a learning rate of 5e-5, AdamW optimizer, and cosine learning rate schedule. We train for 3 epochs with batch size of 16 and gradient accumulation of 3. For DPO, we use a learning rate of 5e-7 and beta parameter of 0.1, training for 5 epochs. All experiments were run on 4 NVIDIA H100 80GB GPUs using bf16 precision and DeepSpeed ZeRO-3.

B.3 Results

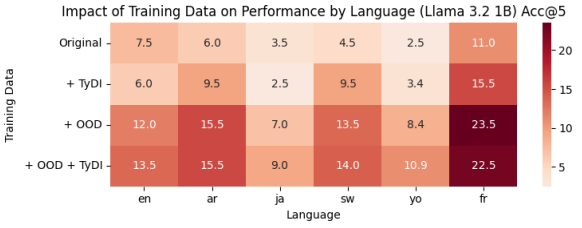

Figure 4: Performance improvement of Llama-3.2-1B over BM25 baseline across languages and metrics.

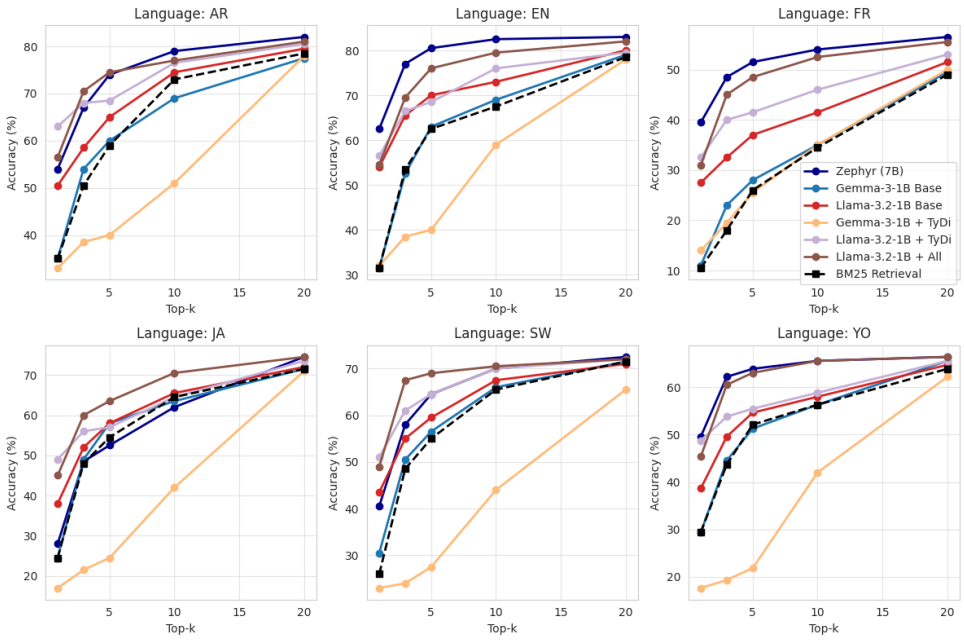

Figure 5: Performance of rerankers across different top-k values by language

C Prompts and Data Example

This section documents the prompt templates used for creating the reasoning-based DPO training datasets.

C.1 Initial Thinking Prompt

The initial prompt used to obtain reasoning processes without revealing golden evidence information:

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

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.
```

C.2 Refinement Prompts

C.2.1 In-Language Thinking Refinement

The prompt used to refine reasoning while maintaining the query language:

```
System: You are RankLLM, an intelligent
↪ assistant that can rank passages
↪ based on their relevancy 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.2.2 Translated Thinking Refinement

The prompt used to refine reasoning with English translation:

```
System: You are RankLLM, an intelligent
↪ assistant that can rank passages
↪ based on their relevancy 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 English while
↪ making clear references to the
↪ passages.

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 ({language}).

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

C.3 DPO Preference Pair Construction

The format for creating DPO preference pairs:

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

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

{contexts}

Search Query: {question}

Rank the passages above based on their
↪ relevance to the search query.
All passages should be included and
↪ listed using identifiers, in
↪ descending order of relevance.
The output format should be [] > [],
↪ e.g., [1] > [2].

Chosen Response:
<think>
{thinking content}
</think>
<answer>
{correct ranking with golden evidence
↪ prioritized}
</answer>

Rejected Response:
<think></think>
<answer>
{incorrect ranking with golden evidence
↪ deprioritized}
</answer>
```

C.4 Complete Sample

Below is a simplified example showing a DPO training pair related to the query "What is stainless steel?" with associated passages (only key passages shown):

```
Query: What is stainless steel?

[8] Stainless steel In metallurgy,
↪ stainless steel, also known as inox
↪ steel or inox from French
```

```
inoxydable (inoxidizable), is a steel
↪  alloy, with highest percentage
↪  contents of iron, chromium,
and nickel, with a minimum of 10.5%
↪  chromium content by mass and a
↪  maximum of 1.2% carbon by mass.

[7] Austenitic stainless steel
↪  Austenitic stainless steel is a
↪  specific type of stainless steel
alloy. Stainless steels may be
↪  classified by their crystalline
↪  structure into four main types:
"austenitic", "ferritic," "martensitic
↪  and duplex."

[2] Stainless steel The Allegheny Ludlum
↪  Corporation worked with Ford on
↪  various concept cars with
stainless steel bodies from the 1930s
↪  through the 1970s to demonstrate the
↪  material's potential.

Chosen Response:
<think>
We prioritize passages that directly
↪  explain what stainless steel is—its
↪  composition and classification.
1. Passage [8] gives the core definition:
↪  it states stainless steel is a steel
↪  alloy with a minimum
    chromium content, precisely
    ↪  answering "what is stainless
    ↪  steel."
2. Passage [7] expands on the definition
↪  by classifying stainless steels into
↪  main types.
3. Passage [2] only illustrates
↪  applications rather than defining
↪  the material.
</think>
<answer>[8] > [7] > [3] > [10] > [2] >
↪  [6] > [1] > [4] > [5] > [9]</answer>

Rejected Response:
<think></think>
<answer>[2] > [7] > [6] > [3] > [4] > [5]
↪  > [1] > [9] > [10] > [8]</answer>
```

This example demonstrates how DPO pairs are structured: the chosen response includes detailed reasoning that correctly prioritizes the definitional passage [8], while the rejected response lacks reasoning and incorrectly ranks an application-focused passage [2] first, placing the core definition passage [8] last.