Evaluating Large Language Models for Cross-Lingual Retrieval

Longfei Zuo[*▲] **Pingjun Hong**[*▲] **Oliver Kraus**[▲]
Barbara Plank[▲🤖] **Robert Litschko**[▲🤖]

▲ MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
🤖 Munich Center for Machine Learning (MCML), Munich, Germany
{zuo.longfei, pingjun.hong}@campus.lmu.de,
{o.kraus2, b.plank, robert.litschko}@lmu.de

Abstract

Multi-stage information retrieval (IR) has become a widely-adopted paradigm in search. While Large Language Models (LLMs) have been extensively evaluated as second-stage reranking models for monolingual IR, a systematic large-scale comparison is still lacking for cross-lingual IR (CLIR). Moreover, while prior work shows that LLM-based rerankers improve CLIR performance, their evaluation setup relies on lexical retrieval with machine translation (MT) for the first stage. This is not only prohibitively expensive but also prone to error propagation across stages. Our evaluation on passage-level and document-level CLIR reveals that further gains can be achieved with multilingual bi-encoders as first-stage retrievers and that the benefits of translation diminishes with stronger reranking models. We further show that pairwise rerankers based on instruction-tuned LLMs perform competitively with listwise rerankers. To the best of our knowledge, we are the first to study the interaction between retrievers and rerankers in two-stage CLIR with LLMs. Our findings reveal that, without MT, current state-of-the-art rerankers fall severely short when directly applied in CLIR.

1 Introduction

Cross-lingual information retrieval (CLIR) aims to retrieve documents written in a different language than the query, which facilitates multilingual access to information. Traditionally, CLIR systems have relied heavily on machine translation (MT) to convert either queries or documents into a shared language, effectively transforming the original cross-lingual task into a monolingual task (Oard, 1998; McCarley, 1999; Zhou et al., 2012; Sun et al., 2020; Lawrie et al., 2022). While this MT-based setting has been the dominant approach, it introduces several practical and methodological limitations: translation increases the query latency; it remains
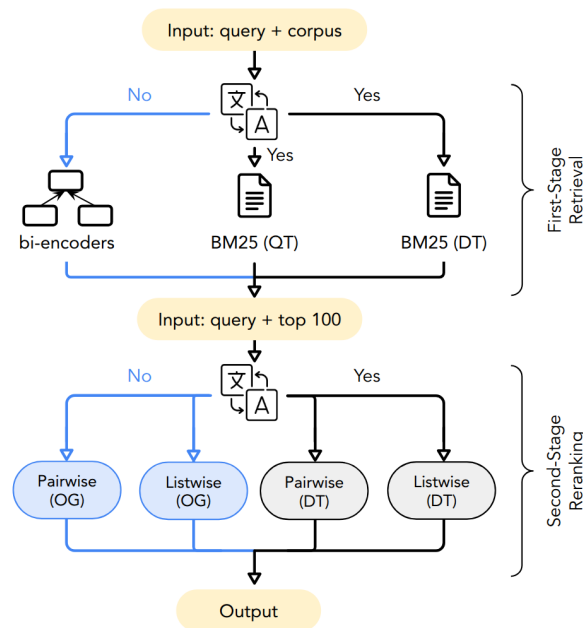

Figure 1: The pipeline used in our experiments. QT: monolingual setup using query translation; DT: monolingual setup using document translation; OG: cross-lingual setup using original documents. Routes marked with blue indicate the cross-lingual information retrieval, where translation is omitted in both the retrieval and reranking stages. Abbreviations used consistently across tables and figures.

unavailable or unreliable for many low-resource languages (Haddow et al., 2022), adversely affecting cross-lingual retrieval when translations contain errors (Litschko et al., 2022a; Guo et al., 2024).

Recent work leveraging large language models (LLMs) in information retrieval has demonstrated promising gains over baseline systems (Ma et al., 2024, 2023), highlighting their capability in ranking tasks. While prior work focuses on monolingual retrieval and reranking, cross-lingual LLM-based retrieval has been understudied. To the best of our knowledge, the only prior works on LLM-based CLIR rely on MT to bridge the language gap in the retrieval stage (Adeyemi et al., 2024a)

11415

and evaluates rerankers on a limited number of languages (Weller et al., 2025b). In our study, we move beyond translation-based setups and conduct a large-scale evaluation of LLMs for cross-lingual retrieval and reranking. Figure 1 illustrates our experimental setup. We compare the performance of state-of-the-art bi-encoders against MT-based lexical retrieval, and quantify the translation gap between rerankers applied on original language documents (OG) versus translated documents (DT). Our work is also the first to compare listwise and pairwise LLM-based rerankers on CLIR.

Our research addresses the following key questions: **RQ1:** How do recent dense multilingual bi-encoders and sparse BM25 differ in their performance as first-stage retrievers for cross-lingual retrieval? **RQ2:** To what extent do LLM rerankers improve retrieval performance when paired with different types of first-stage retrievers, and how does this interaction vary across high-resource and low-resource language settings? **RQ3:** How do pairwise and listwise reranking approaches influence cross-lingual reranking performance? **RQ4:** What is the impact of document length on listwise and pairwise reranking in CLIR?

By addressing these questions, we provide a comprehensive view of LLM rerankers' cross-lingual capabilities and limitations, offering critical guidance for building more effective and adaptable multilingual retrieval solutions. In summary, our main contribution is a systematic evaluation of LLMs for cross-lingual reranking without fully relying on document translation. We release our code and resources to facilitate reproducibility and future research.[1]

2 Related Work

2.1 Multi-Stage Retrieval

The multi-stage retrieval paradigm, widely used in prior work (Nogueira et al., 2019; Ma et al., 2024; Zhuang et al., 2024; Rathee et al., 2025), consists of a fast first-stage retriever followed by one or more reranking stages for improved precision (Nogueira et al., 2019).

Prior studies have shown that first-stage quality can significantly impact reranking performance (Pradeep et al., 2023a,b), and that reranking a smaller, high-quality candidate set (e.g., top-20) can match or exceed reranking larger document pools (e.g., top-100). Lexical retrievers like BM25

(Robertson and Zaragoza, 2009) with document translation pipelines have been widely used in the first stage. More recent approaches employ bi-encoder models to improve candidate quality (Liu et al., 2025b; Lawrie et al., 2024). Our work presents the first large-scale comparative study between LLM-based retrievers and rerankers, across different retrieval paradigms and language pairs.

Recently, many studies have leveraged LLMs as second-stage rerankers due to their strong zero-shot ranking capabilities (Sun et al., 2023; Pradeep et al., 2023a,b; Zhang et al., 2025). Despite their effectiveness in ranking, applying computationally expensive LLM-based rerankers over the full candidate pool is slow and resource-intensive in practice. This makes multi-stage IR crucial to first filter a manageable subset of candidates for reranking.

2.2 Reranking Approaches with LLMs

Listwise Reranking allows LLMs to evaluate multiple candidate documents at once within a single prompt (Ma et al., 2023). Given a query q and a candidate set of documents $D_1, D_2, ..., D_n$, the LLM returns a permutation (i.e., ranked list) based on estimated relevance $D_{\pi(1)}, D_{\pi(2)}, ..., D_{\pi(n)}$.

RankGPT, introduced by Sun et al. (2023), leverages instruction-tuned LLMs for passage reranking and employs a novel prompting-based strategy to enhance listwise reranking performance. However, a key limitation of this work is its reliance on proprietary models. Recent open-source listwise reranking models include RankVicuna (Pradeep et al., 2023a), RankZephyr (Pradeep et al., 2023b) and Rank-without-GPT (Zhang et al., 2025) have been distilled from closed-source models and shown to perform competitive performance on benchmarks like TREC DL (Craswell et al., 2020, 2021) and BEIR (Thakur et al., 2021). However, the performance of LLM-based rerankers on cross-lingual reranking tasks is not well understood, which is a gap that our work aims to address.

Adeyemi et al. (2024a) evaluates RankGPT and RankZephyr on low-resource English-to-African language pairs. However, their evaluation is limited to listwise reranking models with translation-based first-stage retrievers. In this study, we compare their performance against pairwise rerankers and the interaction between rerankers and lexical and dense first-stage retrieval models. Our study involves cross-lingual language pairs on both high-and low-resources languages, as well as language pairs that do not involve English queries.

Pairwise Reranking frames the reranking task as presenting a query along with two candidate documents to a LLM, and the LLM is then prompted to compare the relevance between the candidates in order to select the more relevant one (Qin et al., 2024). In their work, the authors evaluate the proposed pairwise reranking approach on variants of the encoder-decoder FLAN-T5 models (Tay et al., 2023; Chung et al., 2024) on monolingual retrieval benchmarks. We extend this to the cross-lingual domain and evaluate pairwise reranking on two recent multilingual decoder-only LLMs.

Pointwise Reranking is another way to rerank the candidates via relevance generation (Nogueira et al., 2020; Liang et al., 2023; Ma et al., 2024) or query generation strategy (Sachan et al., 2022). Weller et al. (2025b) provide a cross-lingual reranking evaluation of Mistral-7B-Instruct models (Jiang et al., 2023; Weller et al., 2025a) as pointwise rerankers, following the MonoT5 approach (Nogueira et al., 2020). Since research has demonstrated that pairwise methods generally outperform pointwise approaches (Qin et al., 2024; Liu et al., 2025a), we exclusively adopt the pairwise and listwise strategies in our experiments.

3 Evaluation Framework

In the following, we describe our evaluation framework shown in Figure 1. To quantify to what extent gains in the retrieval stage translate to better reranking results, we experiment with translation-based lexical retrieval as well as multilingual bi-encoders. We also compare listwise and pairwise rerankers and quantify the impact of translation.

3.1 Datasets

We select the 2003 portion of the **CLEF** benchmark, due to its well-established cross-lingual test collections comprising both query sets and document corpora (Braschler, 2003). In CLEF 2003 documents consist of newswire articles and cover European languages. Following the experimental setup of Litschko et al. (2022b), we evaluate cross-lingual retrieval experiments across 9 language pairs: EN-{FI, DE, IT, RU}, DE-{FI, IT, RU}, and FI-{IT, RU}, with 60 parallel queries.

Following Adeyemi et al. (2024a), we use the "Test Set A" split of the **CIRAL** benchmark (Adeyemi et al., 2024b) to evaluate CLIR on low-resource African languages. CIRAL is a cross-lingual passage retrieval benchmark where

CLEF 2003				CIRAL (Test Set A)			
Lang.	#Q	#D	Avg. Len.	Lang.	#Q	#D	Avg. Len.
DE	60	295k	284	Ha	80	715k	135
FI	60	55k	256	So	99	827k	126
IT	60	158k	298	Sw	85	949k	127
RU	37	17k	258	Yo	100	82k	168
EN	60	169k	509	–	–	–	–

Table 1: Statistics of the CLEF 2003 and CIRAL (Test Set A) datasets. #Q: number of queries per language; #D: number of documents per language; Avg. Len.: average number of document tokens using a whitespace tokenizer.

queries are written in English and documents written in Hausa (HA), Somali (SO), Swahili (SW) and Yoruba (YO). Dataset statistics are shown in Table 1. CIRAL documents are extracted from African news and blog website and chunked into passages. The average length of CIRAL passages is with 139 white-space-delimited tokens less than that of CLEF documents, which consists of 274 tokens on average. In Section 4.3 we investigate the impact of document lengths on LLM-based reranking.

3.2 Multi-Stage Pipeline: Experimental Setup

Building upon the limitations identified in prior translation-dependent setups, we design a multi-stage retrieval pipeline that evaluates multilingual capabilities more directly. All the evaluation experiments are run on NVIDIA A100 GPUs. A summary of all models used in this study can be found in Table 8.

First-Stage Retrieval. As a lexical baseline retrieval method, we use BM25 via Pyserini (Lin et al., 2021), with batch size of 1 and a single thread. BM25 parameters are set to $k_1 = 0.9$, $b = 0.4$. We index each collection by language and retrieve the top 100 documents for each query. Since lexical retrieval is not well-suited for cross-lingual retrieval, we index documents after translating them to the query language (see Section 3.3).

Furthermore, we chose to evaluate five state-of-the-art bi-encoder models as first-stage retrievers, divided into encoder-only and decoder-only transformer architectures. For each model we retrieve the top 100 documents with a maximum document length of 512, which is a common default value in most models. The evaluated encoder-only models include M3 (Chen et al., 2024), mGTE (Zhang et al., 2024) and multilingual E5 (Wang et al.,

2024). M3 is initialized from XLM-R (Conneau et al., 2020) and supports dense, sparse and multi-vector representations. In our multi-stage pipeline, we utilize the dense representations. mGTE builds on a modernized BERT which is trained from scratch on multilingual datasets. This resulting model is then further optimized for retrieval tasks on a contrastive loss objective. E5 is initialized from XLM-R and then finetuned on a contrastive loss objective. Among the available small, base, and large variants, we use multilingual-e5-large.

The evaluated decoder-only models include RepLLaMA (Ma et al., 2024) and NV-Embed-v2 (Lee et al., 2025). RepLLaMA is initialized from LLaMA-2-7B (Touvron et al., 2023) and then further finetuned for retrieval tasks. As the model is restricted to unidirectional attention, the representation of a document or query is extracted from an appended end-of-sequence token. NV-Embed-v2 is initialized from Mistral-7B (Jiang et al., 2023). During training with a contrastive loss objective, the causal attention mask is removed, enabling bi-directional attention. Embeddings are extracted using a latent attention pooling mechanism rather than relying on a single token embedding. A practical consideration when using decoder-only models is their relatively high storage requirements. Both RepLLaMA and NV-Embed-v2 utilize a 4096-dimensional embedding space, which can lead to significantly larger index sizes and increased memory requirements compared to typical 768- or 1024-dimensional embeddings used in encoder-only models.

Our model selection was informed by their strong performance on existing monolingual benchmarks (Thakur et al., 2021; Craswell et al., 2020, 2021) and the multilingual MMTEB benchmark (Enevoldsen et al., 2025). By evaluating those models on the CLEF and CIRAL datasets, we shed light on their performance on typologically diverse language pairs including low-resource languages.

Listwise Reranking. We include listwise reranking in the second stage of the retrieval pipeline due to its demonstrated efficacy (Sun et al., 2023). Specifically, we employ RankZephyr and RankGPT$_{3.5}$ following Adeyemi et al. (2024a), and additionally incorporate RankGPT$_{4.1}$ into our evaluation. RankZephyr performs well in both monolingual (Pradeep et al., 2023b) and MT-based cross-lingual retrieval (Adeyemi et al., 2024a), but its effectiveness in cross-lingual reranking with-

out MT—where both query and documents are non-English—remains underexplored. We evaluate RankZephyr using the rank_llm framework (Sharifymoghaddam et al., 2025),[2] with a sliding window size of 20, step size of 10, and a max context length of 4096 tokens. The reranking prompt is shown in Figure 3 in Appendix A.

Pairwise Reranking. In our experiments, we utilized the Pairwise Ranking Prompting (PRP) with the bubble-sort-like sliding window strategy (Qin et al., 2024), in which k passes are performed by comparing each document pair from bottom to the top with the initial ranking. Following the prompt design of Qin et al. (2024) (see Figure 4), we perform $k = 10$ passes through the candidate list. We adopt the generation mode, where models output either "Passage A" or "Passage B". The experiments are conducted using two multilingual LLMs: Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Aya-101 (Üstün et al., 2024), due to their strong multilingual capabilities. While Aya-101 has seen all four CIRAL and five CLEF languages, Llama-3.1-8B-Instruct only supports three of those languages (English, German, and Italian).

3.3 Document Translation Setup

To compare cross-lingual and monolingual retrieval performance, we adopt a translation-based setup that converts the task into a monolingual one, isolating the impact of language mismatch.

To maintain consistency with the official translations of the CIRAL dataset (Adeyemi et al., 2024b)—which are generated using the nllb-200-1.3B model (NLLB Team et al., 2022), we also translate the entire CLEF 2003 documents collection for all nine language pairs using the same model. In order to ensure a high translation quality, we use a sentence-level strategy: documents are split into sentences, translated individually, and then concatenated to reconstruct the full document. In a preliminary study, we found that this approach improves the translation accuracy and avoids errors that often arise in document-level translation. In particular, we found that sentence-by-sentence translations reduces word and phrase repetitions, which occurred more often when documents were translated as a whole. All sentence translations use a maximum input sequence length of 128 and batch size of 256 instances. Translations are generated using greedy decoding.

---

[2]https://github.com/castorini/rank_llm

	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG
First-stage retrieval (dense)										
(1a) mGTE	**0.324**	0.302	0.263	0.348	**0.330**	0.307	**0.295**	0.262	0.229	0.296
(1b) RepLLaMA	0.298	0.320	0.305	0.288	0.271	0.286	0.265	0.273	0.160	0.274
(1c) M3	0.309	0.321	0.269	0.298	0.290	0.287	0.225	0.277	0.169	0.272
(1d) E5	0.166	0.223	0.048	0.193	0.211	0.261	0.114	0.279	0.131	0.181
(1e) NV-Embed-v2	0.286	**0.450**	**0.324**	**0.422**	0.148	**0.404**	0.287	**0.342**	**0.244**	**0.323**
(1f) NV-Embed-v2 *(oracle)*	0.499	0.738	0.545	0.683	0.485	0.684	0.539	0.624	0.477	0.586
First-stage retrieval (sparse)										
(2a) BM25-QT	0.309	**0.409**	0.237	0.207	0.267	**0.397**	**0.241**	**0.341**	**0.261**	0.297
(2b) BM25-DT	**0.413**	0.396	**0.255**	**0.485**	**0.301**	0.282	0.216	0.245	0.179	**0.308**
(2c) BM25-DT *(oracle)*	0.613	0.686	0.571	0.716	0.617	0.533	0.481	0.446	0.378	0.560
Listwise Reranking (Retriever: NV-Embed-v2)										
(3a) RankZephyr (OG)	0.351*	0.453	0.340	0.444	0.287*	0.385	0.309	0.300	0.250	0.347
(3b) RankGPT$_{3.5}$ (OG)	0.337	0.438	0.325	0.438	0.297*	0.428	0.305	0.386*	0.250	0.356
(3c) RankGPT$_{4.1}$ (OG)	0.416*	0.526*	**0.417***	0.506*	0.400*	0.501*	**0.410***	0.449*	0.339	0.440
(3d) RankZephyr (DT)	0.402*	0.466	0.375	0.472*	0.347*	0.394	0.318	0.308	0.217	0.367
(3e) RankGPT$_{3.5}$ (DT)	0.383*	0.472	0.308	0.450	0.311*	0.432	0.336	0.375	0.239	0.367
(3f) RankGPT$_{4.1}$ (DT)	**0.433***	**0.557***	0.382	**0.517***	**0.402***	**0.518***	0.363*	**0.472***	**0.341**	**0.443**
Listwise Reranking (Retriever: BM25-DT)										
(4a) RankZephyr (OG)	0.402	0.436	0.304	0.473	0.349	0.350*	0.315*	0.272	0.176	0.342
(4b) RankGPT$_{3.5}$ (OG)	0.405	0.423	0.283	0.479	0.389*	0.338*	0.259	0.288*	0.190	0.339
(4c) RankGPT$_{4.1}$ (OG)	0.480*	0.512*	**0.402***	0.544*	0.465*	**0.421***	0.318*	**0.355***	0.244	0.416
(4d) RankZephyr (DT)	0.461	0.475*	0.333*	0.510	0.425*	0.366*	0.322*	0.230	0.190	0.368
(4e) RankGPT$_{3.5}$ (DT)	0.405	0.467*	0.303*	0.487	0.393*	0.361*	0.278	0.270	0.184	0.350
(4f) RankGPT$_{4.1}$ (DT)	**0.505***	**0.537***	**0.373***	**0.554***	**0.477***	0.413*	**0.341***	0.341*	**0.260**	**0.422**

Table 2: MAP scores of first-stage retriever and listwise reranking results on CLEF 2003, with the best performance for each language pair marked in **Bold**. Gray font indicates the best possible (oracle) reranking performance achievable based on the top-100 documents provided by the first-stage retriever. OG denotes cross-lingual reranking with documents in their original language. QT and DT denote experiments involving query and document translation. *: statistically significant difference to the first-stage retriever (paired t-test, $p < 0.05$).

4 Results and Discussion

4.1 First-Stage Retrieval

Table 2 (rows 1a-1e) presents the results for the first-stage retrievers on the CLEF 2003 dataset. NV-Embed-v2 shows the best average performance (0.323 MAP) with a notable drop in performance for the DE-FI language pair (0.148 MAP), despite having been pre-trained on both languages. mGTE exhibits the second best performance, and is able to outperform NV-Embed-v2 on the two language pairs DE-FI and DE-RU. This is particularly noteworthy, as mGTE is the smallest model in terms of parameter size (see Table 8). We find that E5 shows the weakest overall performance (0.181 MAP). Translation-based lexical retrieval (2a-b) outperforms all bi-encoders except for NV-Embed-v2.

Table 3 (rows 1a-1e) presents the results of the first-stage retrievers on the CIRAL dataset. Here, M3 is the best-performing model, followed by E5, which is notable given its relatively poor performance on CLEF 2003. Both NV-Embed-v2 and mGTE perform substantially worse on CIRAL. We find that language coverage explains this difference only to a certain extent. While NV-Embed-v2 does not support any of the four CIRAL languages, mGTE does support Hausa. Different from our results on CLEF, we find document translation to perform substantially worse. BM25-DT falls behind most bi-encoders, with a notable gap to M3.

These results show that the best-performing first-stage ranker differs considerably based on the chosen dataset (**RQ1**). Meanwhile, it seems likely that main factors for the performance differences are the quality and amount of training data a model has seen. While architectural differences between models (within the classes of encoder-only models and decoder-only models respectively) may play a role, training data composition and quality appears to be a more significant factor. For instance, E5 and M3 share similar architectural foundations, yet their performance diverges substantially on CIRAL.

Compared to prior work that utilized cross-lingual word embeddings (Glavaš et al., 2019; Litschko et al., 2022b; Zhou et al., 2022) and multilingual pre-trained language models (Litschko et al., 2019), we observe that recent bi-encoders achieve superior performance.

	EN-HA	EN-SO	EN-SW	EN-YO	AVG
First-stage retrieval (dense)					
(1a) mGTE	0.252	0.266	0.317	0.339	0.294
(1b) RepLLaMA	0.199	0.192	0.183	0.315	0.222
(1c) E5	0.291	0.278	0.326	0.415	0.327
(1d) NV-Embed-v2	0.136	0.263	0.290	**0.471**	0.290
(1e) M3	**0.388**	**0.351**	**0.402**	0.425	**0.392**
(1f) M3 *(oracle)*	0.744	0.687	0.792	0.793	0.754
First-stage retrieval (sparse)					
(2a) BM25-QT	0.087	0.081	0.130	0.286	0.146
(2b) BM25-DT	**0.214**	**0.246**	**0.233**	**0.445**	**0.285**
(2c) BM25-DT *(oracle)*	0.586	0.561	0.611	0.826	0.646
Listwise Reranking (Retriever: M3)					
(3a) RankZephyr (OG)	0.352	0.302	0.372	0.433	0.365
(3b) RankGPT$_{3.5}$ (OG)	0.419*	0.382*	0.413	0.484*	0.425
(3c) RankGPT$_{4.1}$ (OG)	0.467*	0.453*	0.485*	0.566*	0.493
(3d) RankZephyr (DT)	0.464*	0.454*	0.448*	0.540*	0.477
(3e) RankGPT$_{3.5}$ (DT)	0.439*	0.395*	0.419	0.491*	0.436
(3f) RankGPT$_{4.1}$ (DT)	**0.490***	**0.481***	**0.498***	**0.576***	**0.511**
Listwise Reranking (Retriever: BM25-DT)					
(4a) RankZephyr (OG)	0.260*	0.300*	0.291*	0.439	0.322
(4b) RankGPT$_{3.5}$ (OG)	0.241	0.292	0.256	0.442	0.308
(4c) RankGPT$_{4.1}$ (OG)	0.383*	0.354*	0.361*	0.574*	0.418
(4d) RankZephyr (DT)	0.371*	0.362*	0.365*	0.531*	0.407
(4e) RankGPT$_{3.5}$ (DT)	0.298	0.308	0.307	0.499	0.353
(4f) RankGPT$_{4.1}$ (DT)	**0.397***	**0.378***	**0.406***	**0.584***	**0.441**

Table 3: nDCG@20 scores of first-stage retrievers and listwise rerankers on the CIRAL dataset, with the best performance for each language pair marked in **Bold**. Gray font indicates the best possible (oracle) reranking results achievable based on the top-100 documents provided by the first-stage retriever. *: statistically significant difference to the retriever (paired t-test, $p < 0.05$). Rows 2a and 4b,e are taken from (Adeyemi et al., 2024a).

4.2 Second-Stage Reranking

Listwise Reranking Results. We show the results for listwise rerankers applied on CLEF and CIRAL in Tables 2 and 3. On average, all rerankers manage to improve the results of the input rankings generated by their respective first-stage retrievers. At the level of individual language pairs, we find that only RankGPT$_{4.1}$ improves its input rankings across the board on both datasets. On the CLEF dataset, RankZephyr and RankGPT 3.5 show improved performance over the input rankings generated by NV-Embed-v2 for most language pairs (7-8 out of 9, rows 3a-b), but the improvement is less pronounced when using BM25-DT as the first-stage retriever, with 6-7 language pairs showing improvement (rows 4a-b).

Consistent across both datasets, we find that improved retrieval results translate to better reranking results (**RQ2**). Specifically, as shown in Table 2, we observe that the modest improvement in retrieval results achieved by NV-Embedd-v2 over BM25-DT on the CLEF dataset (+0.015 MAP) translates to relatively small differences in reranking results, ranging from +0.024 for RankGPT$_{3.5}$ (comparing rows 3b and 4b) to +0.05 for RankZephyr (comparing rows 3a and 4a). In contrast, our results on the CIRAL dataset (Table 3) reveal a more substantial improvement in retrieval results, with M3 outperforming BM25-DT by +0.110 MAP. Consequently, the reranking improvements are more pronounced, spanning from +0.043 for RankZephyr (comparing rows 3a and 4a) to +0.117 for RankGPT$_{3.5}$.

The differences between CLEF and CIRAL are likely the result of differences in translation qualities (Adeyemi et al., 2024a) (see also Appendix B). Interestingly, as models achieve an overall stronger performance, the benefits from document translation appears to diminish. On CLEF with NV-Embed-v2 as a first-stage retriever, we observe that improvements for RankZephyr and RankGPT$_{3.5}$ are with +0.026 and +0.011 MAP (comparing rows 3a-b with 3d-e) larger than for RankGPT$_{4.1}$, where it is only +0.003 (comparing 3c with 3f). We observe a similar trend on the CIRAL dataset, where the improvements resulting from document translation for RankZephyr is with +0.112 (3a, 3d) is higher than +0.018 for RankGPT$_{4.1}$ (3c, 3f).

	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG
First-stage retrieval										
NV-Embed-v2	0.286	**0.450**	0.324	0.422	0.148	**0.404**	**0.287**	**0.342**	**0.244**	**0.323**
BM25-DT	**0.413**	0.396	**0.255**	**0.485**	**0.301**	0.282	0.216	0.245	0.179	0.308
Pairwise Reranking (Retriever: NV-Embed-v2)										
Llama-3.1-8B-Instruct (OG)	0.354*	0.474	**0.361**	0.438*	0.282*	0.439*	0.338*	**0.395***	0.254*	0.371
Aya-101 (OG)	0.350*	0.476	0.335	0.430	0.289*	0.436	**0.339**	0.379*	**0.278**	0.368
Llama-3.1-8B-Instruct (DT)	**0.369***	**0.502***	0.358	**0.461***	**0.333***	**0.459***	0.335	0.369	0.276	**0.385**
Aya-101 (DT)	0.337*	0.464	0.326	0.433	0.301*	0.434	0.331	0.352	0.239	0.357
Pairwise Reranking (Retriever: BM25-DT)										
Llama-3.1-8B-Instruct (OG)	0.460*	0.450*	**0.312***	0.492	0.383*	**0.359***	0.269	**0.286***	0.191	0.356
Aya-101 (OG)	0.449*	0.418	0.301	0.498*	0.355*	0.328*	**0.313***	0.274*	**0.234**	0.352
Llama-3.1-8B-Instruct (DT)	**0.479***	**0.472***	0.293	**0.509**	0.390*	0.352*	0.272	0.281*	0.196	**0.360**
Aya-101 (DT)	0.449	0.432*	0.296	0.497	**0.394***	0.336*	0.284*	0.260	0.199	0.350

Table 4: MAP scores of pairwise reranking on CLEF 2003, with the best performance for each language pair marked in **Bold**. *: statistically significant difference to the first-stage retriever (paired t-test, $p < 0.05$).

	EN-HA	EN-SO	EN-SW	EN-YO	AVG
First-stage retrieval					
M3	**0.388**	**0.351**	**0.402**	0.425	**0.392**
BM25-DT	0.214	0.236	0.233	**0.445**	0.282
Pairwise Reranking (Retriever: M3)					
Llama-3.1-8B-Instruct (OG)	0.399	0.360	0.423*	0.453*	0.409
Aya-101 (OG)	0.427*	0.387*	**0.437***	0.483*	0.434
Llama-3.1-8B-Instruct (DT)	0.410*	**0.398***	0.431*	0.474*	0.428
Aya-101 (DT)	**0.431***	0.395*	0.432*	**0.505***	**0.441**
Pairwise Reranking (Retriever: BM25-DT)					
Llama-3.1-8B-Instruct (OG)	0.241*	0.262*	0.280*	0.473*	0.314
Aya-101 (OG)	**0.325***	**0.317***	0.298*	**0.504***	**0.361**
Llama-3.1-8B-Instruct (DT)	0.281*	0.284*	**0.306***	0.503*	0.344
Aya-101 (DT)	0.303*	0.314*	0.305*	0.501*	0.356

Table 5: nDCG@20 scores of pairwise reranking on CIRAL, with the best result for each language pair marked in **Bold**. *: statistically significant difference to the first-stage retriever (paired t-test, $p < 0.05$).

Pairwise Reranking Results. Tables 4 and 5 show the pairwise reranking results on CLEF and CIRAL. On average across all language pairs, we find that all pairwise reranking models improve their input rankings (**RQ3**). The results on CLEF with NV-Embed-v2 as the first-stage retriever[3] show that Llama-3.1-8B-Instruct achieves a performance of 0.371 and 0.385 with and without document translation, outperforming both RankZephyr (0.347 and 0.367) and RankGPT$_{3.5}$ (0.356 and 0.367). The Aya-101 reranker outperforms both listwise models on translated documents.

The results on the CIRAL dataset with M3 as the first-stage retriever show that the better-performing Aya-101 reranker only outperforms RankZephyr and RankGPT$_{3.5}$ when documents remain in their original language. GPT$_{4.1}$ outperforms both pair-wise models. However, it is worth noting that pairwise rerankers are only based on instruction-tuned LLMs and have not been further post-trained on reranking data. This is a key difference to listwise rerankers used in this study, which are either closed-source or have been distilled from closed-source models.

The Glass Ceiling of Reranking. While improvements after second-stage reranking are commonly reported, it remains valuable to examine how closely current methods approximate the optimal ranking. To estimate the possible best result of reranking, we place all relevant documents in the retrieved candidate list at the top positions to simulate the oracle first-stage result (row 1f & 2c in Table 2 & Table 3). The difference between this oracle ranking result and the actual first-stage scores show best possible gains for second-stage reranking.

We define *Potential Reranking Improvements*

---

[3]We focus our discussion on the reranking results based on the better-performing multilingual bi-encoders. In Appendix C we show that further gains are obtainable with hybrid retrieval.

(PRI) and the *Realized* improvement percentage as:

$$\text{PRI} = s^* - s_1 \qquad (1)$$

$$\text{Realized} = \frac{s_2 - s_1}{\text{PRI}} \times 100 \qquad (2)$$

where s_1 denotes the performance of the best dense or sparse first-stage retriever, s^* denotes the corresponding oracle performance achievable at the first stage, and s_2 denotes the best second-stage reranking performance based on the respective s_1.

On CLEF 2003, the PRI is 0.263 for NV-Embed-v2 and 0.252 for BM25 with document translation (DT). Scores on the CIRAL dataset are higher—0.362 for the M3 retriever and 0.361 for BM25 (DT), indicating a larger room for improvement.

In terms of realized improvements, all three rerankers enhance the first-stage results under document translation (DT) settings on both datasets. RankGPT$_{4.1}$ (DT) consistently achieves the best performance, realizing 45.6% and 45.2% of the PRI on the CLEF dataset when reranking NV-Embed-v2 and BM25 (DT) outputs, and 32.9% and 43.2% on the CIRAL dataset when reranking the M3 bi-encoder and BM25 (DT) results. However, the original setting (OG) sometimes fails to outperform the first-stage results (row 3a in Table 3), suggesting that reranking may introduce additional noise in certain cross-lingual settings.

Although document translation can help narrow the performance gap toward the best possible results, the highest realized improvements still fall short of expectations. Both datasets reveal a clear "ceiling effect" in reranking: even state-of-the-art rerankers struggle to approach the upper bound, especially in cross-lingual settings. This highlights a substantial gap between current reranking capabilities and their theoretical potential, suggesting that much of the available improvement remains untapped.

4.3 Impact of Document Length

The passage-level results on CIRAL are noticeably better than the document-level results on CLEF, which motivates us to explore the influence of documents lengths on the reranking performance.

We conduct an ablation study on the CLEF 2003 dataset by varying the maximum number of input tokens per document chunk in both listwise and pairwise setups. This parameter plays a crucial role in balancing information sufficiency and information overload for the reranker.
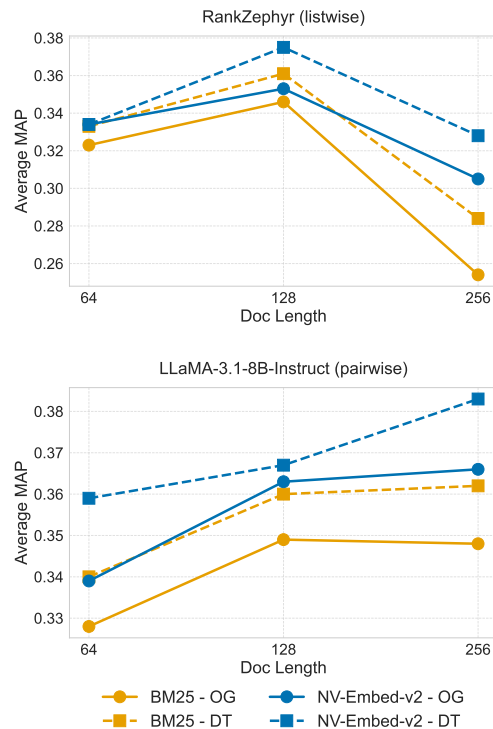

Figure 2: Effect of input document length on reranking across retrievers and reranking approaches. Tokenization is performed using each model's own tokenizer. Results are averaged over all CLEF language pairs.

As shown in Figure 2, RankZephyr reranker performs best at medium input lengths (128 tokens), with a noticeable performance drop at 256 tokens across both sparse (BM25) and dense (NV-Embed-v2) retrievers. This suggests that listwise rerankers are sensitive to overly long inputs, which may dilute useful signals with unnecessary context. In contrast, Llama-based pairwise rerankers show more stable or even improving trends. When using NV-Embed-v2 as the retriever, performance increases steadily from 64 to 256 tokens, indicating that this setup benefits from additional context. However, when using BM25, the gains from longer input saturate or even slightly decline at 256 tokens for DT. This suggests that the pairwise reranker, when paired with high-quality dense retriever, is more robust to longer input spans.

We hypothesize that pairwise reranking is a simpler task and may help capture more nuanced comparison between the two candidates. In this case, longer documents provide additional context, which pairwise rerankers can utilize more effectively (**RQ4**). Overall, these results show that optimal input length is task- and model-dependent and may require careful tuning.

5 Conclusion

In this paper, we conducted a systematic evaluation of LLMs for CLIR, evaluating both passage- and document-level reranking without relying entirely on machine translation. Our results reveal that further cross-lingual reranking gains can be achieved by substituting lexical MT-based retrievers with better-performing multilingual bi-encoder and hybrid retrieval approaches, and that the benefits resulting from document translation diminish with stronger listwise rerankers. We further demonstrate that instruction-tuned pairwise rerankers perform competitively with listwise rerankers like RankZephyr. Finally, we highlight the sensitivity of reranking performance to document length and language resource disparities. Our findings highlight the need for more robust approaches to harness LLMs for cross-lingual reranking without relying on machine translation.

Limitations

While our study provides a comprehensive evaluation of LLMs for CLIR, several limitations remain. First, our evaluation is limited to a fixed set of LLMs, which, while representative, may not represent the full diversity of available open-source or commercial models. Second, although we compare document and query translation strategies, translation quality remains a potential confounder, as noisy translations may affect LLM reranking behavior. Finally, we restrict our experiments to zero-shot reranking without fine-tuning or domain adaptation. This may also underestimate the true potential of rerankers in downstream applications.

Acknowledgments

Ethical considerations. We do not foresee any ethical concerns associated with this work. All analyses were conducted using publicly available datasets and models. No private or sensitive information was used. Additionally, we release our code, prompts, and documentations to support transparency and reproducibility.

Use of AI Assistants. The authors acknowledge the use of ChatGPT exclusively for assistance with grammar, punctuation, and vocabulary refinement, as well as for support with coding-related tasks.

References

Mofetoluwa Adeyemi, Akintunde Oladipo, Ronak Pradeep, and Jimmy Lin. 2024a. 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.

Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, Abdul-Hakeem Omotayo, Idris Abdulmumin, Naome A. Etori, Toyib Babatunde Musa, Samuel Fanijo, Oluwabusayo Olufunke Awoyomi, Saheed Abdullahi Salahudeen, Labaran Adamu Mohammed, Daud Olamide Abolade, Falalu Ibrahim Lawan, Maryam Sabo Abubakar, Ruqayya Nasir Iro, Amina Imam Abubakar, and 4 others. 2024b. Ciral: A test collection for clir evaluations in african languages. SIGIR '24, page 293–302, New York, NY, USA. Association for Computing Machinery.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Martin Braschler. 2003. Clef 2003 - overview of results. volume 3237, pages 44–63.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. *Preprint*, arXiv:2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track. *Preprint*, arXiv:2003.07820.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, and 1 others. 2025. Mmteb: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ping Guo, Yue Hu, Yanan Cao, Yubing Ren, Yunpeng Li, and Heyan Huang. 2024. Query in your tongue: Reinforce large language models with retrievers for cross-lingual search generative experience. In *Proceedings of the ACM Web Conference 2024*, pages 1529–1538.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Dawn Lawrie, Efsun Kayi, Eugene Yang, James Mayfield, and Douglas W. Oard. 2024. Plaid shirttt for large-scale streaming dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2574–2578, New York, NY, USA. Association for Computing Machinery.

Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. 2022. Hc4: A new suite of test collections for ad hoc clir. In *European Conference on Information Retrieval*, pages 351–366. Springer.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1109–1112.

Robert Litschko, Ivan Vulić, and Goran Glavaš. 2022a. Parameter-efficient neural reranking for cross-lingual and multilingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1071–1082, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022b. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.

Junlong Liu, Yue Ma, Ruihui Zhao, Junhao Zheng, Qianli Ma, and Yangyang Kang. 2025a. Listconranker: A contrastive text reranker with listwise encoding. *Preprint*, arXiv:2501.07111.

Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. 2025b. Leveraging passage embeddings for efficient listwise reranking with large language models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4274–4283, New York, NY, USA. Association for Computing Machinery.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2421–2425, New York, NY, USA. Association for Computing Machinery.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

J. Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214, College Park, Maryland, USA. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *Preprint*, arXiv:1910.14424.

Douglas W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 472–483, Langhorne, PA, USA. Springer.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *Preprint*, arXiv:2309.15088.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *Preprint*, 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. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025. Guiding retrieval using llm-based listwise rankers. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I*, page 230–246, Berlin, Heidelberg. Springer-Verlag.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sahel Sharifymoghaddam, Ronak Pradeep, Andre Slavescu, Ryan Nguyen, Andrew Xu, Zijian Chen, Yilin Zhang, Yidi Chen, Jasper Xian, and Jimmy Lin. 2025. Rankllm: A python package for reranking with llms. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3681–3690, New York, NY, USA. Association for Computing Machinery.

Shuo Sun, Suzanna Sia, and Kevin Duh. 2020. CLIReval: Evaluating machine translation as a cross-lingual information retrieval task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 134–141, Online. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 14918–14937, Singapore. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction fine-tuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025a. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.

Orion Weller, Benjamin Chang, Eugene Yang, Mahsa Yarmohammadi, Samuel Barham, Sean MacAvaney,

Arman Cohan, Luca Soldaini, Benjamin Van Durme, and Dawn Lawrie. 2025b. mfollowir: A multilingual benchmark for instruction following in retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part II*, page 295–310, Berlin, Heidelberg. Springer-Verlag.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Crystina Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. 2025. Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part II*, page 233–247, Berlin, Heidelberg. Springer-Verlag.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Dong Zhou, Wei Qu, Lin Li, Mingdong Tang, and Aimin Yang. 2022. Neural topic-enhanced cross-lingual word embeddings for clir. *Information Sciences*, 608:809–824.

Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Comput. Surv.*, 45(1).

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 38–47. ACM.

A Prompt Template for Listwise and Pairwise Reranking

Figures 3 and 4 show the prompt templates used in our experiments. For listwise reranking, we adopt the prompt template from the original RankZephyr implementation (Pradeep et al., 2023b). For the pairwise reranking approach, we use the prompt introduced in (Qin et al., 2024) and implementation provided by Zhuang et al. (2024).

```
LISTWISE RERANKING PROMPT

<|systems|>
You are RankLLM, an intelligent assistant that
can rank passages based on their relevancy to
the query.
<|user|>
I will provide you with {num} passages, each
indicated by a numerical identifier []. Rank
the passages based on their relevance to the
search query: {query}.

[1] {passage 1}
[2] {passage 2}
...
[{num}] {passage {num}}

Search Query: {query}.

Rank  the  {num}  passages  above  based  on
their relevance to the search query. All the
passages  should  be  included  and  listed  using
identifiers,  in  descending  order  of  relevance.
The output format should be [] > [], e.g., [4]
> [2].  Only respond with the ranking results,
do not say any word or explain.
<|assistant|>

Model Generation: [9] > [4] > [20] > ... > [13]
```

Figure 3: Listwise reranking prompt template used for LLM-based reranking.

```
PAIRWISE RERANKING PROMPT

Given a query {query}, which of the following
two passages is more relevant to the query?

Passage A: {document₁}
Passage B: {document₂}

Output Passage A or Passage B:
```

Figure 4: Pairwise reranking prompt template used for LLM-based reranking.

B Query and Document Translation

We compare the impact of query translation (QT) and document translation (DT) strategies across the two benchmarks. As shown in Table 2 and Table 3,

While both QT and DT are considered in the first stage, we only use DT to construct the can-

didate pool for the second stage due to its better performance. Notably, the performance gap between QT and DT is much larger on CIRAL than on CLEF in the first stage. QT is less effective on CIRAL, as translating queries into low-resource African languages often produces short or low-quality queries that struggle to match the document content. By contrast, QT performs better on CLEF, where queries are translated into higher-resourced languages with more reliable translation and greater lexical overlap.

For the reranking stage, applying DT on CIRAL has a much more pronounced effect than on CLEF, resulting in a large performance gap between DT and OG. We attribute this to two factors: (1) translating documents into English transforms the reranking task into a noisy EN–EN format, which aligns more closely with the LLM's training distribution, and (2) CIRAL documents are relatively short, leading to higher translation accuracy and less noise compared to longer texts.

In contrast, the OG setting on CLEF is less problematic. The languages included are better supported by existing LLMs, and the queries themselves are typically longer and more descriptive (both title and description are included in queries), making them easier to interpret and translate. As a result, the performance gap between OG and DT is smaller.

C Hybrid Retrieval Experiments

In addition to our BM25 and bi-encoder retrieval experiments, we also evaluate a hybrid retrieval approach in which LLMs rerank fused top-100 rankings using reciprocal rank fusion (Cormack et al., 2009).[4] In the following, we limit our analysis to open-source LLMs. The retrieval and reranking results are shown in Tables 6 and 7.

Oracle results. With regard to the best possible reranking performance achievable, we find mixed results. On CLEF, we notice that hybrid retrieval improves the reranking potential (i.e., oracle score) to 0.611 MAP, while both reranking the input rankings of NV-Embedd-v2 and BM25-DT can at best yield MAP values of 0.586 and 0.560. However, on CIRAL we find that the best possible reranking results of the hybrid model falls below the perfor-

---

[4]The union of the top-100 document sets of BM25 and bi-encoder is larger than 100. For a fair comparison, we also limit the number of documents of the fused ranking to 100 documents.

	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG
First-stage retrieval										
NV-Embed-v2	0.286	0.450	0.324	0.422	0.148	**0.404**	0.287	0.342	**0.244**	0.323
BM25-DT	**0.413**	0.396	0.255	0.485	0.301	0.282	0.216	0.245	0.179	0.308
Hybrid	0.407	**0.480**	**0.336**	**0.516**	**0.372**	0.406	**0.303**	**0.344**	0.221	**0.376**
NV-Embed-v2 *(oracle)*	0.499	0.738	0.545	0.683	0.485	0.684	0.539	0.624	0.477	0.586
BM25-DT *(oracle)*	0.613	0.686	0.571	0.716	0.617	0.533	0.481	0.446	0.378	0.560
Hybrid *(oracle)*	0.648	0.776	0.657	0.758	0.665	0.702	0.614	0.625	0.505	0.661
Listwise Reranking (Retriever: hybrid)										
RankZephyr (OG)	0.399	0.479	**0.393**	0.509	0.386	0.398	0.359	0.320	**0.211**	0.384
RankZephyr (DT)	0.468*	0.489	0.393	0.528	0.462*	0.440	0.387*	0.319	0.206	0.410
Pairwise Reranking (Retriever: hybrid)										
Llama-3.1-8B-Instruct (OG)	**0.447***	0.502	**0.407***	0.513	0.406	0.451*	0.367	0.389*	**0.301***	0.420
Llama-3.1-8B-Instruct (DT)	0.463*	0.508*	0.381	**0.535***	**0.436***	**0.471***	0.339	0.388*	0.294	**0.424**
Aya-101 (OG)	0.443*	**0.509**	0.363	0.522	0.417*	0.459*	**0.373***	**0.397***	0.298	0.420
Aya-101 (DT)	0.436*	0.501	0.374	0.524	0.421*	0.457*	0.369*	0.359	0.247	0.410

Table 6: MAP scores of hybrid retriever and listwise reranking results on CLEF 2003, with the best performance for each language pair and retrieval stage marked in **Bold**. *: statistically significant difference to the first-stage retriever (paired t-test, $p < 0.05$).

	EN-HA	EN-SO	EN-SW	EN-YO	AVG
First-stage retrieval					
M3	**0.388**	0.351	**0.402**	0.425	0.392
BM25-DT	0.214	0.236	0.233	0.445	0.282
Hybrid	0.382	**0.358**	0.377	**0.497**	**0.403**
M3 *(oracle)*	0.744	0.687	0.792	0.793	0.754
BM25-DT *(oracle)*	0.586	0.561	0.611	0.826	0.646
Hybrid *(oracle)*	0.586	0.687	0.792	0.793	0.715
Listwise Reranking (Retriever: hybrid)					
RankZephyr (OG)	0.359	0.336	0.390	0.477	0.390
RankZephyr (DT)	**0.474***	**0.462***	**0.474***	**0.571***	**0.495**
Pairwise Reranking (Retriever: hybrid)					
Llama-3.1-8B-Instruct (OG)	0.397*	0.367	0.416*	0.517*	0.424
Llama-3.1-8B-Instruct (DT)	0.417*	**0.405***	0.418*	**0.540***	0.445
Aya-101 (OG)	0.455*	0.402*	**0.427***	0.536*	**0.455**
Aya-101 (DT)	**0.456***	0.404*	0.422*	0.535*	0.454

Table 7: nDCG@20 scores of hybrid retrieval and reranking on the CIRAL dataset, with the best performance for each language pair and retrieval stage marked in **Bold**. *: statistically significant difference to the first-stage retriever (paired t-test, $p < 0.05$).

mance of the M3 retriever. This may be explained by the fact that the gap between the bi-encoder and lexical retriever is much larger on CIRAL (0.392 vs. 0.282) than on CLEF (0.323 vs. 0.308). Interestingly, despite the large performance gap between both prerankers on CIRAL, the hybrid model still brings slight performance improvements (with a MAP score of 0.403).

Listwise reranking results. On listwise reranking on CLEF, we find that the performance of RankZephyr improves from 0.342 (see Table 2) to 0.384 without translation (OG), and from 0.368 to 0.410 when documents are translated (DT). Similar gains can be seen on CIRAL (see Table 3), where the MAP scores improve from 0.327 to 0.390

with original language documents (OG), and from 0.407 to 0.495 with translated documents. These results are consistent with our results presented in Section 4 and show that improvements in retrieval translate to improvements in reranking.

Pairwise reranking results. Different from the listwise reranking results, we find that the improvements resulting from document translation diminish for pairwise rerankers. Both Llama-3.1-8B-Instruct and Aya-101 perform substantially better compared to reranking only the results of NV-Embedd-v2 and M3 for CLEF and CIRAL (see Tables 4 and 5). On CLEF, the best performance is achieved by Llama-3.1-8B (DT), whereas RankZephyr (DT) performs best on CIRAL.

Model	Parameters	#Lang.	Unsupported Languages	Emb. Dim.
mGTE (Zhang et al., 2024)	305M	75	HA	768
RepLlama (Ma et al., 2024)	7B	1	FI, DE, IT, RU, HA, SW, SO, YO	4096
M3 (Chen et al., 2024)	560M	173	HA	1024
E5 (Wang et al., 2024)	560M	94	YO	1024
NV-Embedd-v2 (Lee et al., 2025)	7.85B	1	FI, DE, IT, RU, HA, SW, SO, YO	4096
RankZephyr (Pradeep et al., 2023b)	7.24B	1	FI, DE, IT, RU, HA, SW, SO, YO	-
RankGPT$_{3.5}$ (Sun et al., 2023)	-	-	-	-
RankGPT$_{4.1}$ (Sun et al., 2023)	-	-	-	-
Aya-101 (Üstün et al., 2024)	12.9B	101	None	-
Llama-3.1-8B-Instruct (Touvron et al., 2023)	8.03B	8	FI, RU, HA, SW, SO, YO	-

Table 8: Retriever and reranker model information based on HuggingFace model cards (Wolf et al., 2020). For each model, we list the number of parameters, number languages it was trained on, and which of the CIRAL and CLEF languages are not supported. For bi-encoders we also report the embedding size.

D Model Information

Information about the retriever and reranker models used in our study is summarized in Table 8. For retrievers, RepLlama is fine-tuned from Llama-2-7B (Touvron et al., 2023) using LoRA (Hu et al., 2021) on the MS MARCO Passage Ranking (Bajaj et al., 2018) training split for one epoch, while E5 is initialized from XLM-RoBERTa-large (Conneau et al., 2020) and trained on a mixture of multilingual datasets covering 100 languages. NV-Embed-v2 model is built upon decoder-only Mistral-7B-v0.1 (Jiang et al., 2023).

For rerankers, RankZephyr is based on the Zephyr-7B-ß model (Tunstall et al., 2023). We use the RankZephyr Full version, which is distilled from RankGPT$_{3.5}$ (100K training queries) and RankGPT$_4$ (5K queries). The training corpus is limited to monolingual English data.