Zero-Shot Cross-Lingual Reranking with Large Language Models for Low-Resource Languages

Mofetoluwa Adeyemi, Akintunde Oladipo, Ronak Pradeep, Jimmy Lin

David R. Cheriton School of Computer Science
University of Waterloo

{moadeyem, aooladipo, rpradeep, jimmylin}@uwaterloo.ca

Abstract

Large language models (LLMs) as listwise rerankers have shown impressive zero-shot capabilities in various passage ranking tasks. Despite their success, there is still a gap in existing literature on their effectiveness in reranking low-resource languages. To address this, we investigate how LLMs function as listwise rerankers in cross-lingual information retrieval (CLIR) systems with queries in English and passages in four African languages: Hausa, Somali, Swahili, and Yoruba. We analyze and compare the effectiveness of monolingual reranking using either query or document translations. We also evaluate the effectiveness of LLMs when leveraging their *own* generated translations. To grasp the general picture, we examine the effectiveness of multiple LLMs—the proprietary models RankGPT_4 and $\text{RankGPT}_{3.5}$, along with the open-source model RankZephyr. While the document translation setting, i.e., both queries and documents are in English, leads to the best reranking effectiveness, our results indicate that for specific LLMs, reranking in the African language setting achieves competitive effectiveness with the cross-lingual setting, and even performs better when using the LLM's own translations.

1 Introduction

Several studies have shown that large language models (LLMs) excel in various NLP tasks (Zhou et al., 2022; Zhu et al., 2023; Wang et al., 2023). In text ranking, LLMs have been used effectively as retrievers (Ma et al., 2023a) and in both pointwise and listwise reranking. In reranking, models may generate an ordered list directly (Sun et al., 2023; Ma et al., 2023b; Pradeep et al., 2023a; Tamber et al., 2023) or sort based on token probabilities (Ma et al., 2023b). The large context size of LLMs makes listwise approaches particularly attractive because the model attends to multiple documents to produce a relative ordering.

Cross-lingual retrieval aims to provide information in a language different from that of the search query. This is especially relevant when the required information is not available or prevalent in the query's language, as is the case for most low-resource languages. Previous work has examined sparse and multilingual dense retrieval models in cross-lingual settings for these languages (Zhang et al., 2023b; Ogundepo et al., 2022). However, studies on the effectiveness of LLMs as cross-lingual retrievers or rerankers for low-resource languages are few to non-existent.

In this study, we examine the effectiveness of proprietary and open-source models for listwise reranking in low-resource African languages. Our investigation is guided by the following research questions: (1) How well do LLMs fare as listwise rerankers for low-resource languages? (2) How effectively do LLMs perform listwise reranking in cross-lingual scenarios compared to monolingual (English or low-resource language) scenarios? (3) When we leverage translation, is reranking more effective when translation uses the same LLM used for zero-shot reranking?

We answer these questions through an extensive investigation of the effectiveness of RankGPT (Sun et al., 2023) and RankZephyr (Pradeep et al., 2023b) in cross-lingual and monolingual retrieval settings. We use CIRAL (Adeyemi et al., 2023), a cross-lingual information retrieval dataset covering four African languages with queries in English and passages in African languages, and construct monolingual retrieval scenarios through document and query translations.

Our results show that cross-lingual reranking with these LLMs is generally more effective compared to reranking in the African languages, underscoring that they are better tuned to English than low-resource languages. Across all languages, we achieve our best results when reranking entirely in English using retrieval results obtained by doc-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 650–656

ument translation. In this setting, we see up to 7 points improvement in nDCG@20 over cross-lingual reranking using RankGPT$_4$, and up to 9 points over reranking in African languages. We specifically notice improvements with RankGPT$_4$ when using its query translations for reranking in African languages.

2 Background and Related Work

Given a corpus $C = \{D_1, D_2, ..., D_n\}$ and a query q, information retrieval (IR) systems aim to return the k most relevant documents. Modern IR pipelines typically feature a multi-stage architecture in which a first-stage *retriever* returns a list of candidate documents that a *reranker* reorders for improved quality (Asadi and Lin, 2013; Nogueira et al., 2019; Zhuang et al., 2023).

More recently, the effectiveness of decoder models as rerankers (dubbed "prompt decoders") has been explored in some depth. Researchers have fine-tuned GPT-like models in the standard contrastive learning framework (Neelakantan et al., 2022; Muennighoff, 2022; Zhang et al., 2023a) and studied different approaches to reranking using both open-source LLMs and proprietary GPT models. Sun et al. (2023) evaluated the effectiveness of OpenAI models on multiple IR benchmarks using permutation generation approaches, while Ma et al. (2023b) demonstrate the effectiveness of GPT-3 as a zero-shot listwise reranker and the superiority of listwise over pointwise approaches.

While these papers focus on reranking with LLMs, they only cover two African languages—Swahili and Yoruba. For both languages, GPT-3 improves over BM25 significantly but still falls behind supervised reranking baselines. In this work, we examine the effectiveness of these LLMs as components of IR systems for African languages. Specifically, we study the effectiveness of open-source and proprietary LLMs as listwise rerankers for four African languages (Hausa, Somali, Swahili, and Yoruba) using the CIRAL cross-lingual IR test collection (Adeyemi et al., 2023).

To be more precise, cross-lingual information retrieval (CLIR) is a variant of the standard retrieval task in which the queries q_i are in a different language from the documents in the corpus C. Popular approaches to CLIR include query translation, document translation, and language-independent representations (Lin et al., 2023). As the focus of this work is on the effectiveness of LLMs as listwise rerankers in cross-lingual settings, we primarily explore document and query translation approaches in this study.

Input Prompt:

```
SYSTEM
You are RankGPT, 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 number identifier [].
Rank the passages based on their relevance
to the 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.
The passages should be listed in descending
order using identifiers. The most relevant
passages should be listed first. The output
format should be [] > [], e.g., [1] > [2].
Only respond with the ranking results, do not
say any word or explain.
```

Model Completion:

```
[10] > [4] > [5] > [6] ... [12]
```

Figure 1: Prompt design and sample of model completion adopted for listwise reranking with the LLMs.

3 Methods

Listwise Reranking. In listwise reranking, LLMs compare and attribute relevance over multiple documents in a single prompt. As this approach has been proven to be more effective than pointwise and pairwise reranking (Ma et al., 2023b; Pradeep et al., 2023a), we solely employ listwise reranking in this work. For each query q, a list of provided documents $D_1, ..., D_n$ is reranked by the LLM, where n denotes the number of documents that are inserted into the prompt.

Prompt Design. We adopt RankGPT's (Sun et al., 2023) listwise prompt design as modified by Pradeep et al. (2023a). The input prompt and generated completion are presented in Figure 1.

LLM Zero-Shot Translations. We examine the effectiveness of LLMs in using their translations in crossing the language barrier. For a given LLM, we generate zero-shot translations of queries from English to African languages and implement reranking with the LLM using its translations. With this approach, we are able to examine the ranking effec-

Input Prompt:

```
Query: {query}
Translate this query to {African language}.
Only return the translation, don't say any
other word.
```

Model Completion:

```
{Translated query}
```

Figure 2: Prompt design and model completion for zero-shot query translations with the LLMs.

tiveness of the LLM solely in African languages, and examine the correlation between its translation quality and reranking. The prompt design for generating the query translation is shown in Figure 2.

4 Experimental Setup

Models. We implement zero-shot reranking for African languages with three models. These include proprietary reranking LLMs: RankGPT_4 and $\text{RankGPT}_{3.5}$, using the `gpt-4` and `gpt-3.5-turbo` models, respectively, from Azure's OpenAI API. To examine the effectiveness of open-source LLMs, we rerank with RankZephyr (Pradeep et al., 2023b), an open-source reranking LLM obtained by instruction-fine-tuning Zephyr_β (Tunstall et al., 2023) to achieve competitive effectiveness with RankGPT models.

Baselines. We compare the reranking effectiveness of the LLMs using already established models as baselines. Our baselines include two cross-encoder models, the multilingual T5 (mT5) (Xue et al., 2021) and AfrimT5 (Adelani et al., 2022), which is mT5 with continued pre-training on African corpora. The mT5[1] and AfrimT5[2] rerankers were obtained from fine-tuning the base versions of both models on the MS MARCO passage collection (Bajaj et al., 2016) for 100k iterations, with a batch size of 128.

Test Collection. Models are evaluated on CIRAL (Adeyemi et al., 2023), a CLIR test collection consisting of four African languages: Hausa, Somali, Swahili, and Yoruba. Queries in CIRAL are natural language factoid questions in English while passages are in the respective African languages. Each language comprises between 80 and 100 queries, and evaluations are done using the pooled judgments obtained from CIRAL's passage retrieval task.[3] We also make use of CIRAL's translated passage collection[4] in our document translation scenario. The test collection's documents were translated from the African languages to English using the NLLB machine translation model (Costa-jussà et al., 2022).

We report nDCG@20 scores following the test collection's standard, and MRR@100.

Configurations. First-stage retrieval uses BM25 (Robertson and Zaragoza, 2009) in the open-source Pyserini toolkit (Lin et al., 2021). We use whitespace tokenization for passages in native languages and the default English tokenizer for the translated passages. Our BM25 retrieval is implemented using document (BM25-DT) and query (BM25-QT) translations. For BM25-QT, queries are translated with Google Machine Translation (GMT).

We rerank the top 100 passages retrieved by BM25 using the sliding window technique by Sun et al. (2023) with a window of 20 and a stride of 10. Experiments were conducted using the RankLLM toolkit.[5] We use a context size of 4,096 tokens for $\text{RankGPT}_{3.5}$ and RankZephyr, and 8,192 tokens for RankGPT_4. These context sizes are also maintained for the zero-shot LLM translation experiments. For each model, translation is performed over three iterations and we vary the model's temperatures from 0 to 0.6 to allow variation in the translations. Translations are only obtained for the GPT models since RankZephyr is suited only for reranking. Reranking results are reported over a single run, except with the LLM translations where we take the Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) of results from the three iterations.

5 Results and Discussion

5.1 Cross-Lingual vs. Monolingual Reranking

Table 1 compares results for cross-lingual reranking using CIRAL's queries and passages unmodified, and also the English reranking scenario. Row (1) reports scores for the two first-stage retrievers, BM25 with query translation (BM25-QT) and document translation (BM25-DT). Cross-lingual reranking scores for the different LLMs are presented in Row (2), and we employ BM25-DT for first-stage retrieval given it is more effective.

[1] https://huggingface.co/castorini/mt5-base-ft-msmarco

[2] https://huggingface.co/castorini/afrimt5-base-ft-msmarco

[3] https://ciralproject.github.io/

[4] https://huggingface.co/datasets/CIRAL/ciral-corpus#translated-dataset

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

	Source		nDCG@20					MRR@100				
	Prev.	top-k	ha	so	sw	yo	Avg	ha	so	sw	yo	Avg
(1a) BM25-QT	None	$\|C\|$	0.0870	0.0813	0.1302	0.2864	0.1462	0.1942	0.1495	0.3209	0.4434	0.2770
(1b) BM25-DT	None	$\|C\|$	0.2142	0.2461	0.2327	0.4451	0.2845	0.4009	0.4050	0.4426	0.5904	0.4597
Cross-lingual Reranking: English queries, passages in African languages												
(2a) RankGPT$_4$	BM25-DT	100	**0.3577**	**0.3159**	**0.3029**	**0.5070**	**0.3709**	**0.7006**	**0.5613**	**0.6378**	**0.7364**	**0.6590**
(2b) RankGPT$_{3.5}$	BM25-DT	100	0.2413	0.2919	0.2562	0.4416	0.3078	0.5125	0.5151	0.5615	0.5932	0.5456
(2c) RankZephyr	BM25-DT	100	0.2741	0.2941	0.2953	0.4459	0.3274	0.4917	0.5195	0.5884	0.6311	0.5577
(2d) mT5	BM25-DT	100	0.3876	0.3757	0.3778	0.5604	**0.4254**	0.6381	0.6294	0.6855	0.6938	**0.6617**
(2e) AfrimT5	BM25-DT	100	0.3911	0.3530	0.3655	0.5510	0.4152	0.6463	0.5998	0.6888	0.6903	0.6563
English Reranking: English queries, English passages												
(3a) RankGPT$_4$	BM25-DT	100	**0.3967**	**0.3819**	**0.3756**	**0.5753**	**0.4324**	**0.7042**	0.6125	**0.7112**	**0.7523**	**0.6951**
(3b) RankGPT$_{3.5}$	BM25-DT	100	0.2980	0.3080	0.3074	0.4985	0.3530	0.5702	0.5373	0.6241	0.7306	0.6156
(3c) RankZephyr	BM25-DT	100	0.3686	0.3630	0.3678	0.5275	0.4067	0.6431	**0.6210**	0.6995	0.7169	0.6701
(3d) mT5	BM25-DT	100	0.3644	0.3877	0.3587	0.5489	0.4149	0.5916	0.6104	0.6335	0.6732	0.6272
(3e) AfrimT5	BM25-DT	100	0.3748	0.3663	0.3591	0.5499	0.4125	0.6333	0.5521	0.6160	0.6983	0.6249

Table 1: Comparison of Cross-lingual and English reranking results. The cross-lingual scenario uses CIRAL's English queries and African language passages while English reranking crosses the language barrier with English translations of the passages.

Scores for reranking in English are reported in Row (3), and results show this to be the more effective scenario across the LLMs and languages. However, the cross-encoder T5 baselines have better reranking effectiveness in the cross-lingual scenario.

Improved reranking effectiveness with English translations is expected, given that LLMs, despite being multilingual, are more attuned to English. The results obtained from reranking solely with African languages further probe the effectiveness of LLMs in low-resource language scenarios. We report scores using query translations in Table 2, with BM25-DT also as the first-stage retriever for a fair comparison. In comparing results from the query translation scenario to the cross-lingual results in Row (2) of Table 1, we generally observe better effectiveness with cross-lingual. However, RankGPT$_4$ obtains higher scores for Somali, Swahili, and Yoruba in the African language scenario, especially with its query translations, comparing Rows (2a) in Table 1 and 2.

5.2 LLM Reranking Effectiveness

We compare the effectiveness of the different LLMs across the reranking scenarios. RankGPT$_4$ generally achieves better reranking among the 3 LLMs, as presented in Tables 1 and 2. In the cross-lingual and English reranking scenarios, the open-source LLM RankZephyr (Pradeep et al., 2023b) achieves better reranking scores in comparison with RankGPT$_{3.5}$ as reported in Rows (*b) and (*c) in Table 1. RankZephyr also achieves comparable scores with RankGPT$_4$ in the English reranking scenario, and even a higher MRR for Somali as reported in Row (3c) of Table 1. These results establish the growing effectiveness of open-source LLMs for language tasks considering the limited availability of proprietary LLMs, but with room for improvement in low-resource languages.

In comparing the reranking effectiveness of LLMs with that of the baseline models, scores vary depending on the scenario and specific LLM. Reranking scores of the cross-encoder T5 baselines are reported in Rows (*d) and (*e) of Tables 1 and 2. As seen in Rows (2d) and (2e) of Table 1, the cross-encoder multilingual T5 baselines achieve higher reranking scores compared to all three LLMs. However, RankGPT$_4$ outperforms both baselines in the English reranking scenario and using its query translations in the African language reranking scenario. We can attribute the higher effectiveness of the baselines to being fine-tuned for reranking as compared to the LLMs where reranking is carried out in a zero-shot fashion.

5.3 LLM Translations and Reranking

Given that RankGPT$_4$ achieves better reranking effectiveness using its query translations in the monolingual setting, we further examine the effectiveness of this scenario. Row (2) in Table 2 reports results using LLMs translations, and we compare these to results obtained using translations from GMT. Compared to results obtained with GMT translations, RankGPT$_4$ does achieve better monolingual reranking effectiveness in the African language using its query translations. RankGPT$_{3.5}$ on the other hand achieves less competitive scores on average using its query translations when com-

	Source		nDCG@20					MRR@100				
	Prev.	**top-k**	**ha**	**so**	**sw**	**yo**	**Avg**	**ha**	**so**	**sw**	**yo**	**Avg**
(1) BM25-DT	None	$\|C\|$	0.2142	0.2461	0.2327	0.4451	0.2845	0.4009	0.4050	0.4426	0.5904	0.4597
LLM Query Translations: Queries and passages in African languages												
(2a) RankGPT_4	BM25-DT	100	0.3458	**0.3487**	**0.3559**	**0.4834**	**0.3835**	0.6293	0.4253	**0.6961**	**0.6551**	0.6015
(2b) $\text{RankGPT}_{3.5}$	BM25-DT	100	0.2370	0.2773	0.2802	0.4462	0.3102	0.4651	0.4756	0.5314	0.6115	0.5209
GMT Query Translations: Queries and passages in African languages												
(3a) RankGPT_4	BM25-DT	100	**0.3523**	0.3086	0.3086	0.4712	0.3602	**0.6800**	**0.5154**	0.6252	0.6545	**0.6188**
(3b) $\text{RankGPT}_{3.5}$	BM25-DT	100	0.2479	0.2816	0.2761	0.4361	0.3104	0.4996	0.4741	0.5647	0.5505	0.5222
(3c) RankZephyr	BM25-DT	100	0.2515	0.2520	0.2556	0.4114	0.2926	0.4573	0.4407	0.5460	0.5690	0.5033
(3d) mT5	BM25-DT	100	0.3395	0.3305	0.3412	0.4963	**0.3769**	0.5313	0.5105	0.5551	0.6574	0.5636
(3e) AfrimT5	BM25-DT	100	0.3559	0.3335	0.3428	0.4620	0.3736	0.5863	0.5195	0.6028	0.5886	0.5743

Table 2: Reranking in African languages using query translations and passages in the African language. BM25-DT is used as first stage. Query translations are done using the LLMs, and we compare effectiveness with GMT translations.

Model	**ha**	**so**	**sw**	**yo**	**avg**
GPT_4	21.8	7.4	43.8	16.0	22.3
$\text{GPT}_{3.5}$	7.1	1.8	42.4	6.6	14.5
GMT	45.3	17.9	85.9	36.7	46.5

Table 3: Evaluation of the LLMs' query translation quality using the BLEU metric. Scores reported are the average over three translation iterations.

pared to translations from the GMT model, with the exception of Yoruba where it has much higher scores using its translations.

Considering the effect of translation quality on reranking, we evaluate the LLMs' translations and report results in Table 3. Evaluation is done against CIRAL's human query translations using the BLEU metric. We observe better translations with GPT_4 compared to $\text{GPT}_{3.5}$, with GMT achieving the best quality. However, RankGPT_4 still performs better using its query translations, indicating a correlation in the model's understanding of the African languages.

6 Conclusion

In this work, we evaluate zero-shot cross-lingual reranking with large language models (LLMs) on African languages. Our suite covered three forms of LLM-based reranking: RankGPT_4, $\text{RankGPT}_{3.5}$ and RankZephyr. Using the listwise reranking method, our results demonstrate that reranking in English via translation is the most optimal. We examine the effectiveness of LLMs in reranking for low-resource languages in the cross-lingual and African language monolingual scenarios and find that LLMs have comparable effectiveness in both scenarios but with better results in cross-lingual. In the process, we also establish that good translations obtained from the LLMs do improve their reranking effectiveness in the African language reranking scenario as discovered with RankGPT_4.

Additionally, while open-source models showcase slightly lower effectiveness than RankGPT_4, they still largely improve over other proprietary models like $\text{RankGPT}_{3.5}$, an important step towards the development of effective listwise rerankers for low-resource languages.

7 Limitations

While we provide valuable insights into the application of LLMs for reranking tasks in low-resource settings, our work is not without limitations. One constraint is the reliance on translations for achieving good reranking effectiveness, which inherently introduces dependencies on the quality of translation models and their compatibility with the target languages. Additionally, the scope of languages and models evaluated in this study, covering only a *small* spectrum of African languages and a mix of proprietary and open-source LLMs, remains limited in the broader context of low-resource language research.

Future research directions could address these limitations by exploring a wider array of low-resource languages and incorporating more diverse LLMs, including those specifically trained or fine-tuned on low-resource language datasets. Investigating alternative reranking pipelines that reduce reliance on translation or enhance the multilingual capabilities of LLMs directly could also offer new avenues for improving retrieval effectiveness in low-resource language settings.

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council

(NSERC) of Canada and Huawei Technologies Canada. Thanks to Microsoft for providing access to OpenAI LLMs on Azure via the Accelerating Foundation Models Research program. We also thank the anonymous reviewers for their constructive suggestions.

References

David Adelani et al. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070.

Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, and Jimmy Lin. 2023. CIRAL at FIRE 2023: Cross-Lingual Information Retrieval for African Languages. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

Nima Asadi and Jimmy Lin. 2013. Effectiveness/Efficiency Tradeoffs for Candidate Generation in Multi-stage Retrieval Architectures. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *ArXiv*, abs/1611.09268.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Marta R. Costa-jussà et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv*, abs/2207.04672.

Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, Nandan Thakur, Jheng-Hong Yang, and Xinyu Crystina Zhang. 2023. Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval. *ArXiv*, abs/2304.01019.

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 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023a. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. *ArXiv*, abs/2310.08319.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023b. Zero-Shot Listwise Document Reranking with a Large Language Model. *ArXiv*, abs/2305.02156.

Niklas Muennighoff. 2022. SGPT: GPT Sentence Embeddings for Semantic Search. *ArXiv*, abs/2202.08904.

Arvind Neelakantan et al. 2022. Text and Code Embeddings by Contrastive Pre-Training. *ArXiv*, abs/2201.10005.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. *ArXiv*, abs/1910.14424.

Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. 2022. AfriCLIRMatrix: Enabling Cross-Lingual Information Retrieval for African Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8721–8728.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *ArXiv*, abs/2309.15088.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *ArXiv*, abs/2312.02724.

Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

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.

Manveer Singh Tamber, Ronak Pradeep, and Jimmy Lin. 2023. Scaling down, LiTting up: Efficient zero-shot listwise reranking with seq2seq encoder-decoder models. *ArXiv*, abs/2312.16098.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct Distillation of LM Alignment. *ArXiv*, abs/2310.16944.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named Entity Recognition via Large Language Models. *ArXiv*, abs/2304.10428.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023a. Language Models are Universal Embedders. *ArXiv*, abs/2310.08232.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023b. Toward Best Practices for Training Multilingual Dense Retrieval Models. *TOIS*, 42(2):1–33.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large Language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *ArXiv*, abs/2304.04675.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023. Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels. *ArXiv*, abs/2310.14122.