Relevant for the Right Reasons? Investigating Lexical Biases in Zero-Shot and Instruction-Tuned Rerankers

Yuchen Mao[♠*] **Barbara Plank**[▲🤖] **Robert Litschko**[▲🤖]
[♠] Department of Language Science and Technology, Saarland University, Germany
[▲] MaiNLP, Center for Information and Language Processing (CIS), LMU Munich, Germany
[🤖] Munich Center for Machine Learning (MCML), Munich, Germany
yuchmao@lst.uni-saarland.de {b.plank, robert.litschko}@lmu.de

Abstract

Large Language Models (LLMs) show strong potential for reranking documents in information retrieval (IR), but training with monolingual data often leads to monolingual overfitting and lexical bias, limiting generalization in cross-lingual IR (CLIR). To overcome these issues, we investigate instruction-tuning LLaMA-3.1-8B-Instruct on English and multilingual code-switched data, and evaluate on mMARCO and XQuAD-R. Results show that instruction-tuning on code-switched data substantially improves CLIR performance, while monolingual tuning remains more effective for monolingual reranking. We introduce a novel measure to analyze the relationship between lexical overlap and reranking performance, showing that the two factors are correlated. We finally conduct a causal analysis using counterfactual examples, where we evaluate whether rewriting passages that share overlapping keywords with the query causes models to change their relevance predictions. Overall, we find that code-switching serves as an effective and lightweight strategy to improve cross-lingual generalization in LLM-based re-ranking, while our analyses show that lexical overlap remains a major factor that can mislead reranking models.

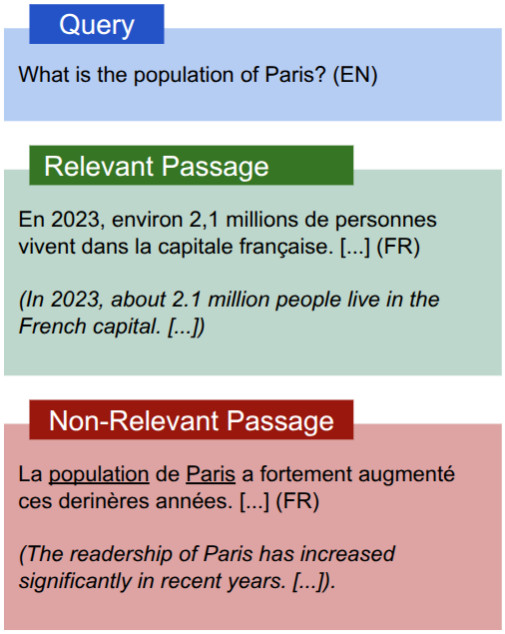

Figure 1: The first passage is semantically relevant to the query but shares no lexical overlap. In contrast, the second passage contains lexical overlap with the query terms "population" and "Paris" but is topically unrelated. Lexically biased LLM rerankers may incorrectly favor the non-relevant passage.

1 Introduction

Large Language Models (LLMs) such as LLaMA-3 (Dubey et al., 2024), GPT-4 (OpenAI et al., 2024), Gemini (Team et al., 2025), and Mistral (Jiang et al., 2023) have shown strong performance across a wide range of NLP tasks. In information retrieval (IR), which aims to return relevant documents from large text collections given a user query, recent advances have led to growing interest growing interest in leveraging LLMs as rerankers. In particular, LLMs have been explored as pointwise (Zhuang et al., 2023; Sun et al., 2023), pairwise (Qin et al., 2024), or list-wise rerankers (Tang et al., 2024; Chen et al., 2025; Parry et al., 2024; Ma et al., 2023) under prompt-based inference settings, where the model refines the order of documents within the initial retrieval set. In parallel, many LLMs have demonstrated their capability to process and generate text in multiple languages (Dang et al., 2024). This progress has further opened new possibilities for the use of LLMs in cross-lingual information retrieval (CLIR), where queries and documents are written in different languages. Recent work has begun to systematically evaluate the performance of LLMs in cross-lingual retrieval settings. For example, Zuo et al. (2025) benchmarked a wide range of LLM rerankers under translated and non-translated CLIR scenarios, analyzing list-

[*] Work done while at LMU Munich.

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 468–482

wise and pairwise strategies as well as the interaction between first-stage retrievers and second-stage rerankers. However, they do not investigate *how* LLMs make relevance judgments.

Understanding whether LLMs determine relevance for the right reasons (i.e., semantic relevance), or whether they are biased towards lexical matches (i.e., shortcuts) is crucial for *equitable information access* and ensuring the trustworthiness of LLM-based retrieval systems (Litschko et al., 2023b). Biases in cross-lingual retrieval settings have been well-studied in the context of multilingual pre-trained language models (mPLMs). Prior work includes studies on, e.g., language biases in mPLM-based bi-encoders (Laosaengpha et al., 2025; Huang et al., 2024; Yang et al., 2024; Roy et al., 2020). Our work is closest to (Litschko et al., 2023a), who study zero-shot cross-lingual transfer of mPLM-based cross-encoders, where models trained on English data have been found to exhibit poor transfer performance to cross-lingual reranking tasks. The authors show that this monolingual overfitting can be mitigated by training on code-switched data instead, which naturally reduces the lexical overlap between queries and documents. However, it remains unclear whether LLMs exhibit similar lexical biases when used as rerankers, and whether instruction-tuning those models on code-switched training data also leads to similar improvements. Figure 1 illustrates this issue for a single pairwise cross-lingual reranking step: the model incorrectly prefers a lexically overlapping but semantically irrelevant passage, suggesting that relevance judgments may not always reflect genuine semantic understanding. This motivates our central question: Are LLM-based reranker outputs relevant for the right reasons?

To address this, we investigate whether LLM-based rerankers are affected by monolingual overfitting and lexical bias, and how instruction tuning strategies change this behavior. Specifically, we compare direct zero-shot reranking (without further training) against instruction-tuning on monolingual English data, multilingual code-switched data and target language-pair data on both MoIR and CLIR. In addition to our reranking experiments, we also characterize the lexical bias through a correlation and causal analysis. Our main contributions are:

- We show that instruction-tuning pair-wise rerankers on code-switched data improves their cross-lingual reranking performance. However, unlike mPLM-based cross-encoders, these gains come at the cost of a worse monolingual reranking performance.
- We introduce two overlap-sensitive metrics, ALOD and AP–LOD correlation, to quantify the link between lexical overlap and reranking quality. Our results show that the two are **positively correlated**. However, this correlation is weak, underpinning that lexical overlap are only one of multiple factors (and biases) influencing what rerankers deem relevant.
- We evaluate the **causal relationship** between lexical overlaps and reranking performance. Specifically, we construct counterfactual examples from previously incorrectly classified instances (see Figure 1) and investigate whether removing lexical overlap by rewriting the passage causes rerankers to recover from incorrect predictions.

2 Related Work

Shortcut Learning in Language Models. Several recent studies have investigated shortcut learning behavior in LLMs, where models rely on superficial features in the input, such as lexical overlap or specific keywords, instead of performing genuine semantic reasoning. Du et al. (2021) focus on BERT-based models and show that these models tend to favor shortcut tokens early in training. Tang et al. (2023) found that LLMs often rely on shallow cues from prompts during in-context learning, rather than understanding the task itself. Sun et al. (2024) showed that instruction tuning and reinforcement learning with human feedback can increase shortcut learning in LLMs across tasks such as reasoning. Yuan et al. (2024) provided a systematic evaluation of shortcut biases, including lexical overlap, in prompt-based inference. Hagstrom et al. (2025) found that LM-based rerankers can be misled by lexical similarities, often favoring candidates with high surface overlap over semantically more relevant passages on English-only retrieval tasks. The study shows that these biases can lead to significant drops in model accuracy. Taken together, these studies suggest that shortcut learning remains a major challenge for LLMs.

However, these works do not explore how shortcut bias behavior changes when LLMs are fine-tuned for monolingual and cross-lingual pairwise reranking. We fill this gap and study shortcut learning behavior in prompt-based reranking tasks, and

especially regarding the model's sensitivity to lexical overlap.

Bias in Multilingual and Cross-lingual Contexts. Gao et al. (2025) analyzed LLMs' cross-lingual context retrieval ability on cross-lingual machine reading comprehension (xMRC). They observed a significant performance gap between monolingual and cross-lingual settings, and propose a two-phase explanation: the model first encodes the question and then retrieves the answer. This highlighted that performance degradation in xMRC is not solely due to output generation but is rooted in earlier stages of processing. While their work identifies where in the model such limitations arise, it does not fully clarify whether relevance decisions are based on semantic features or surface-level lexical shortcuts, which is the focus of our work.

Beyond retrieval tasks, cross-lingual inconsistencies have also been observed across a range of tasks involving semantic understanding, reasoning, and prompt sensitivity. Wang et al. (2024) found that multilingual models fail to achieve balanced performance across languages, with significant disparities depending on the language used. Lai et al. (2023) showed that ChatGPT performs better in English than in other languages, particularly on tasks requiring complex reasoning, with performance gaps especially notable in lower-resource languages. Furthermore, Etxaniz et al. (2024) showed that LLMs often fail to realize their full multilingual potential when prompted in non-English languages, highlighting an implicit preference for English in reasoning processes.

However, these studies do not examine how different instruction tuning strategies affect LLM performance in monolingual versus cross-lingual information retrieval tasks, nor do they address whether such biases differ under different training conditions. Our work aims to fill this gap by systematically comparing reranking behavior under monolingual and code-switched instruction tuning setups.

3 Methodology

We conduct three different types of analyses: First we investigate how well LLM rerankers instruction-tuned on English data generalize to other monolingual reranking (MoIR) and cross-lingual reranking (CLIR) tasks, or whether they suffer from monolingual overfitting (Section 3.1). We then propose a measure that captures the correlation between lexical overlap and reranking performance (Section 3.2). Finally, we introduce an evaluation protocol that facilitates a causal analysis of the impact of lexical overlap on reranking performance at the instance-level (Section 3.3).

3.1 Pair-wise Reranking

This pipeline consists of three steps. (1) We convert monolingual and code-switched training sets into a unified instruction–output format. (2) We fine-tune the base LLM under different language settings. (3) We evaluate the tuned models using pairwise prompting with a sliding window, following (Qin et al., 2024). Each prompting unit is defined as $u(q, d_1, d_2)$, where q is a query and d_1, d_2 are two candidate documents. To obtain the full ranking, we apply a sliding window approach: starting from a randomly shuffled ranking, we iteratively traverse the list in reverse order, comparing and potentially swapping adjacent document pairs (stride = 1) based on model judgments. For each query, we repeat this process ten times to obtain the final top-10 reranked results. The prompting template we use is provided in Appendix A.

For reranking evaluation, we report the results using the metric MRR@10 implemented in the ir_measures package (MacAvaney et al., 2022). To further understand the impact of superficial token overlap, we introduce two complementary metrics to analyze the model's reliance on lexical overlap, as discussed next.

3.2 Correlation Analysis

The first metric captures the average lexical overlap difference (**ALOD**) in lexical overlap between relevant and irrelevant documents (lexical overlap difference, **LOD**) for a given query. For a query q, we compute:

$$\mathrm{LOD}_q = \frac{1}{|D_q^+|} \sum_{d \in D_q^+} \mathrm{Overlap}(q, d) - \frac{1}{|D_q^-|} \sum_{d \in D_q^-} \mathrm{Overlap}(q, d)$$

where D_q^+ and D_q^- denote the sets of relevant and irrelevant documents for query q, respectively, and $\mathrm{Overlap}(q, d)$ denotes the lexical overlap score between q and d, computed as the number of shared tokens (after normalization and stopword removal). We opted for LOD_q instead of simple lexical overlap to ignore shared non-keyword tokens that can be found in both relevant and non-relevant documents. On the dataset-level, **ALOD** is the average

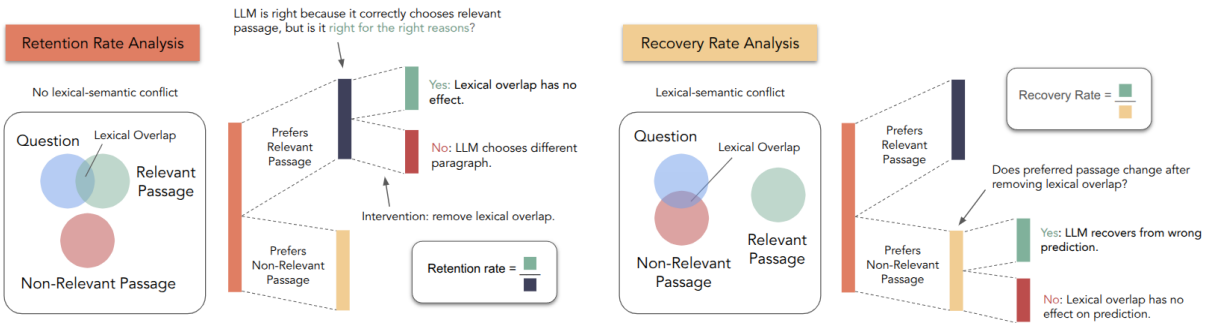

Figure 2: We conduct two types of causal analyses to understand how models determine relevance. **Left**: We use retention rate to measure the extent to which LLMs still correctly prefer the relevant passages (true positives) after removing lexical overlapping keywords. A high retention rate indicates a low lexical bias. **Right**: We use the recovery rate to measure the extent to which errors made by LLMs are due to being misguided by lexical biases. A high recovery rates indicate high lexical bias.

of LOD_q over all queries:

$$\text{ALOD} = \frac{1}{|Q|} \sum_{q \in Q} \text{LOD}(q)$$

ALOD quantifies the degree to which lexical bias can be present in monolingual and cross-lingual ranking datasets. This baseline version of ALOD provides a simple and transparent measure of lexical bias. In our reranking setup, we compute LOD based on the negative documents found in the top-k input ranking. To assess its robustness, we additionally experimented with alternative pre-processing settings, including stopword removal, lemmatization, and subword tokenization, as well as varying the number of negative documents per query. We find that while these variations changed the absolute ALOD values, the relative trends remained consistent with the comparisons above. This confirms that the ALOD metric is robust to preprocessing choices and evaluation settings. Detailed results are provided in Appendix F, Table 11.

The second is **AP–LOD Correlation**, which measures the Spearman correlation (Zar, 2005) between the average precision (**AP**) (Harman, 1992) of each query and its LOD. This correlation captures the alignment between lexical bias and actual ranking performance.

These metrics are applied to both MoIR and CLIR outputs to compare lexical reliance across language settings. Higher ALOD scores indicate a larger potential for models falling back to a lexically bias, while a high AP–LOD correlation shows that this is strongly related to the reranking performance of different models.

3.3 Causal Analysis

Inspired by counterfactual explanations (Verma et al., 2024) and adversarial robustness studies on multilingual embedding models (Michail et al., 2025), we design two types of counterfactual experiments to test to what extent lexical overlap impacts a model's notion of relevance. Here, we conduct our analysis at the instance level, where each sample consists of a query, a relevant passage, and a non-relevant passage. We initially evaluate LLMs on queries that share tokens with the relevant and non-relevant passages, respectively. We then repeat our experiments with perturbed passages, where we remove the lexical overlap (intervention) and measure how it affects the model performance. The original dataset and perturbations are automatically generated with GPT-5 (OpenAI, 2025) (see prompt templates in Appendix D). Using synthetic examples allows us to disentangle the effects of semantic relevance and lexical bias in a controlled way. As shown in Figure 2, we investigate model predictions from two complementary perspectives:

Right for the right reasons? Here, we construct instances where the relevant passage shares keyword tokens with the question, while the non-relevant passage is lexically distinct from the query (Figure 6). We focus on instances where LLMs correctly prefer the relevant passage (henceforth, True Positives – TP), and test if removing lexical overlap (intervention; Figure 8) causes LLMs to change their preferred passages. We compute the **retention rate** as the fraction of TP instances where the intervention has no impact. High retention rates indicate a low lexical bias, where models prefer the relevant passage for the right (semantic) reasons.

Wrong because of lexical overlaps? In this experiment, we generate samples where only the non-relevant passages share keyword tokens with the query, while the relevant ones do not (Figure 7). Here we ask the question whether errors made by LLMs are due to an over-reliance on lexical cues. We focus on errors where models incorrectly prefer the non-relevant passage (henceforth, False Positives – FP), and compute the **recovery rate**, defined as the proportion of FP errors that are corrected after keyword overlap cues are removed (intervention; Figure 9). A high recovery rate indicates a high lexical bias and captures the extent to which models judge documents as relevant for the wrong reasons, specifically caused by lexical overlap.

To ensure the correctness of lexical–semantic conditions (see Figure 2), we prompted GPT-5 multiple times, each time generating 20 candidate examples for a given condition, and accumulated 240 candidates per condition before applying filtering criteria. In the lexical-semantic conflict dataset, the irrelevant passages share lexical tokens with the query while relevant passage do not, and vice versa for the other dataset.After filtering, we obtained 204 conflict and 200 non-conflict instances for the retention and recovery rate analyses.

4 Experimental Setup

4.1 Model and Baselines

We use Llama-3.1-8B-Instruct (Dubey et al., 2024) as the base model for zero-shot reranking (**`Zero-shot`** model) and instruction tuning. During both training and inference, we adopt the official LLaMA-3.1 chat template as the prompting format.[1] We compare this model against different models instruction-tuned on code-switched queries (**`EN-XX-tuned`**) or code-switched queries and documents (**`XX-XX-tuned`**). Hyperparameters are provided in Appendix B. An example prompt using the chat format is shown in Appendix C.

To assess the impact of instruction-tuning on lexical overlap behavior, we construct several variants of Llama-3.1-8B-Instruct.

The **`EN-EN-tuned`** model is instruction-tuned on English monolingual data and serves as our primary baseline. This setup corresponds to the standard zero-shot cross-lingual transfer setting (Lauscher et al., 2020).

The **`EN-XX-tuned`** and **`XX-XX-tuned`** models are tuned on code-switched queries, and on both code-switched queries and documents, respectively, to evaluate the effect of multilingual and mixed-language supervision.

As an upper bound, we include the **`Fine-tuned`** model, which is directly instruction-tuned on the target language pairs and evaluated on corresponding reranking tasks. While this provides a performance reference, it is important to note that this baseline often cannot be reached in practice due to limited language coverage of existing machine translation systems and lack of available instruction-tuning training data.

4.2 Datasets

Following Litschko et al. (2023a), we use the multilingual MS MARCO dataset (mMARCO) dataset (Bonifacio et al., 2022) for model training and evaluation.[2] For instruction tuning, we reuse the public training data provided by Litschko et al. (2023a) in the HuggingFace repository,[3] which was originally derived from the `Train Triple Small` set in the multilingual mMARCO dataset.[4] For the code-switched training data. Specifically, we use the multilingual code-switched data (EN-XX and XX-XX code-switched data) with a translation probability $p = 0.5$. From this pool, for each language pair, we use a sampl of 1 million instances for training.

For evaluation, we construct a reduced version of the dataset, denoted as `top100.dev` from the original `top1000.dev` set provided by mMARCO by keeping all qrels-marked relevant documents from `top1000.dev`, discarding queries without them, and randomly sampling non-relevant ones to obtain 100 documents per query. For each query, the order of its 100 documents is randomly shuffled.

To validate whether other findings generalize to other datasets, we also include the XQuAD-R (Roy et al., 2020) dataset. Here, too, we construct for each query input rankings consisting of top-100 documents. Following the original setup in Roy et al. (2020), we train the model using

[1] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/

[2] The mMARCO dataset includes 14 languages with varying levels of resource availability and writing systems: Arabic (AR), Chinese (ZH), Dutch (NL), English (EN), French (FR), German (DE), Hindi (HI), Indonesian (ID), Italian (IT), Japanese (JA), Portuguese (PT), Russian (RU), Spanish (ES), and Vietnamese (VI).

[3] https://huggingface.co/datasets/rlitschk/csclir/tree/main

[4] https://github.com/spacemanidol/MSMARCO/blob/master/Ranking/README.md

	EN	DE	AR	IT	RU	AVG	AVG_{X-X}	Δ^{ZS}	Δ^{ZS}_{X-X}
Zero-shot	51.14	38.73	29.57	38.58	36.42	38.89	35.83	-	-
EN-EN-tuned	**72.41**	**62.08**	**55.34**	**62.30**	**61.57**	**62.74**	**60.32**	**+23.85**	**+24.50**
EN-XX-tuned	70.72	60.27	45.45	59.31	57.93	58.74	55.74	+19.85	+19.92
XX-XX-tuned	70.50	60.86	45.64	61.46	57.79	59.25	56.44	+20.36	+20.62
Fine-tuned	-	64.79	57.91	65.28	62.35	64.55	62.58	+25.66	+26.76

Table 1: MoIR: Monolingual re-ranking results on mMARCO language pairs in terms of MRR@10. Results are reported per language and averaged in two ways: (1) **AVG** includes all monolingual pairs, and (2) $\mathbf{AVG_{X-X}}$ excludes EN–EN. Δ^{ZS}: Improvement over the zero-shot baseline, computed based on AVG. Δ^{ZS}_{X-X}: Improvement over the zero-shot baseline, computed based on AVG_{X-X}. **Bold**: The best performance for each language (excluding the fine-tuned baseline model).

	EN-DE	EN-IT	EN-AR	DE-IT	DE-RU	AR-IT	AR-RU	AVG	AVG_{X-X}	Δ^{ZS}	Δ^{ZS}_{X-X}
Zero-shot	46.14	46.22	40.64	35.72	34.19	25.52	25.78	36.31	30.30	-	-
EN-EN-tuned	62.87	62.90	**52.61**	54.00	53.43	40.96	**42.42**	52.74	47.71	+16.43	+17.40
EN-XX-tuned	64.21	63.55	51.17	53.31	52.09	33.31	34.31	50.28	43.25	+13.96	+12.95
XX-XX-tuned	**64.63**	**64.51**	51.70	**56.32**	**54.39**	**41.34**	41.01	**53.42**	**48.26**	**+17.10**	**+17.96**
Fine-tuned	66.21	66.54	59.38	61.09	60.01	53.53	52.53	59.90	56.79	+23.58	+26.49

Table 2: CLIR: Cross-lingual re-ranking results on mMARCO in terms of MRR@10.

the English-only SQuAD dataset and its machine-translated versions generated via Google Translate (Wu et al., 2016). For the code-switched version of the SQuAD-based training data, we implement the same code-switching method with a translation probability $p = 0.5$ following the approach in (Litschko et al., 2023a).

We evaluate our pairwise rerankers on a mix of high- and low-resource languages, covering diverse scripts and language families. Specifically, for mMARCO, we include monolingual re-ranking in English (EN), German (DE), Arabic (AR), Italian (IT), and cross-lingual re-ranking in EN-{DE, AR, IT}, DE-{IT, RU} and AR-IT. For XQuAD-R, we select three languages for MoIR (EN, DE, AR) and evaluate CLIR on the following language pairs: EN-{DE, AR}, DE-RU.

We conduct the lexical overlap perturbation experiment on the mMARCO dataset, focusing on four language pairs that include English: one monolingual pair (EN-EN) and three cross-lingual pairs (EN-DE, EN-IT, and EN-AR).

5 Results and Discussion

In the following, we first measure the performance gap of LLMs in monolingual reranking (MoIR) and cross-lingual reranking (CLIR). We specifically investigate how well different instruction-tuning strategies impact the generalization performance. We then validate our findings on XQuAD-R.

5.1 Overall Reranking Results

Cross-task Generalization Performance. Tables 1 and 2 report the MRR@10 scores on five MoIR and seven CLIR language pairs on mMARCO under different training conditions. We also report the average across all language-pairs and language-pairs that do not involve English. Across all settings, models perform better on MoIR than CLIR. For example, the Zero-shot model achieves an average MRR@10 of 0.389 for MoIR versus 0.363 for CLIR. When language-pairs involving are excluded, the gap widens (MoIR: 0.358, CLIR: 0.303). After EN–EN tuning, MoIR reaches 0.627, while CLIR falls behind with a MRR@10 of 0.527. This gap widens to 0.12 if language-pairs involving English are excluded. Similar patterns hold for EN-XX-tuned (0.587 vs. 0.503, gap: 0.084) and XX-XX-tuned (0.593 vs. 0.534, gap: 0.059).

These results show that monolingual reranking is generally easier for LLMs than cross-lingual reranking. This is expected since rerankers do not have to rely on interlingual semantics. Even under instruction-tuning on code-switched data, which improves overall CLIR performance, the gap between MoIR and CLIR remains substantial. This could be due to mismatching vocabularies, where models can rely less on lexical shortcuts in CLIR compared to MoIR. We will explore their correlation and causal relationships further in Section 6.

Instruction-Tuning on English versus Code-Switched Data. Across all MoIR and CLIR lan-

	MoIR					CLIR				
	EN	DE	AR	RU	AVG	EN-DE	EN-AR	DE-RU	AR-RU	AVG
`Zero-shot`	96.87	94.93	90.50	94.08	94.09	96.15	92.99	87.06	84.44	90.16
`EN-EN-tuned`	97.81	96.27	**93.89**	96.37	96.08	96.83	92.12	94.35	88.21	92.88
`EN-XX-tuned`	97.73	**96.56**	93.67	**96.85**	**96.20**	**97.34**	**95.01**	**95.57**	**87.23**	**93.79**
`XX-XX-tuned`	**98.47**	96.44	92.74	96.75	96.10	97.24	93.19	95.39	87.17	93.25
`Fine-tuned`	97.82	96.49	94.05	96.15	96.12	96.77	94.92	95.45	92.61	94.94

Table 3: MoIR and CLIR re-ranking results on XQuAD-R in terms of MRR@10.

guage pairs, models fine-tuned on the target language pair (`Fine-tuned`) consistently achieves the best performance, while the `Zero-shot` model performs the worst. This is expected because fine-tuning on cross-lingual data allows the model to jointly align interlingual semantics and learn ranking-specific features.

For all MoIR language pairs, the `EN-EN-tuned` model consistently outperform models trained on code-switched data, even on non-English monolingual pairs. For example, it achieves a MRR@10 score of 0.724 on English, outperforming both the `EN-XX-tuned` model (0.707) and `XX-XX-tuned` model (0.705). Similarly on Russian, where it yields a performance 0.616 MRR@10, also outperforming the `EN-XX-tuned` (0.579) and `XX-XX-tuned` (0.578) variants. We find a consistent trend of LLM rerankers performing worst on monolingual reranking in Arabic and Russian reranking tasks.

In contrast, CLIR performance generally benefits more from instruction-tuning on code-switched data. For example, on EN–DE, the `XX-XX-tuned` model attains 0.646, outperforming `EN-EN-tuned` (0.629). On AR–IT, it scores 0.413, slightly above EN–EN tuning (0.410). The only exceptions are EN–AR and AR–RU, where `EN-EN-tuned` reranker remains superior (0.526 vs. 0.512/0.517, and 0.424 vs. 0.343/0.410). The cross-lingual reranking performance tends to improve when the question and answer passage languages are typologically more similar. While the `XX-XX-tuned` model performs well on EN–DE (0.646) and EN–IT (0.645), it yields worse results on AR–IT (0.413) and AR–RU (0.410).

Overall, our results indicate that instruction-tuning on code-switched data improves cross-lingual reranking performance. However, contrary to findings reported on mBERT-based cross-encoders (Litschko et al., 2023a), we find a performance trade-off, where code-switching training data improves CLIR at the expense of performance drops in MoIR. We hypothesize that this is related to the syntactic coherence, or the lack thereof in code-switched data,[5] of passages provided in context. The results also reveal a clear English-centric bias: in MoIR, all rerankers achieve the strongest performance on reranking English passages; in CLIR, rerankers perform better on language-pairs involving English queries. Excluding CLIR language-pairs involving English leads to a sharp drops in CLIR performance, ranging from -0.031 (`Fine-tuned` reranker) to -0.070 MRR@10 (`EN-XX-tuned` reranker). The consistently weaker results on Arabic and Russian, and cross-lingual language-pairs involving those languages, suggests that LLM rerankers struggle to bridge the script gap (Chari et al., 2025).

5.2 Evaluation on XQuAD-R

Table 3 presents the reranking performance of models evaluated on XQuAD-R after instruction tuning on the (code-switched) English SQuAD dataset. The results generally follow similar trends to those observed on mMARCO, especially regarding the benefits of code-switching CLIR data. Consistent with our results on mMARCO, we find on CLIR that instruction-tuning variants improve upon the `Zero-shot` model (0.902), `EN-XX-tuned` (0.938) and `XX-XX-tuned` models (0.933) outperform the `EN-EN-tuned` model (0.930), and the model `Fine-tuned` on target the language-pairs performs best (0.949). While the results are overall much higher than those reported on mMARCO, we find that the improvements on CLIR from code switching are much smaller. Taken together, this suggests that the benefits of reducing the lexical overlap in instruction-tuning diminish as the reranking task become easier. In the rest of this paper we focus our analysis on the mMARCO dataset.

[5]The dictionaries used for code switching were induced from nearest cross-lingual neighbors in a multilingual word embedding space. Because of this, there is no guarantee that substituted words belong to the same word class.

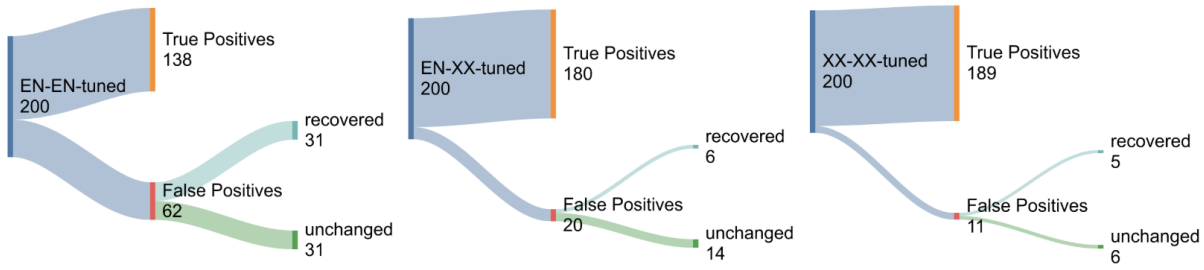

Figure 3: Results of the recovery rate analysis: Sankey diagrams illustrating model decisions on synthetic pairwise reranking experiments before and after perturbation. Non-relevant passages share overlapping keyword tokens with queries, while relevant passages have no overlap. Results are shown for the `EN-EN-tuned`, `EN-XX-tuned`, and `XX-XX-tuned` models. The `Zero-shot` model (not shown) obtained perfect results without any false positives.

	MoIR		CLIR	
	ALOD	$\rho^{\text{AP-LOD}}$	ALOD	$\rho^{\text{AP-LOD}}$
`Zero-shot`	0.90	22.10	0.20	10.49
`EN-EN-tuned`	0.90	28.16	0.20	18.19
`EN-XX-tuned`	0.90	25.61	0.20	14.46
`XX-XX-tuned`	0.90	21.85	0.20	14.81

Table 4: ALOD: Average lexical overlap difference computed separately for MoIR and CLIR on mMARCO. $\rho^{\text{AP-LOD}}$: Spearman correlation (in %) between the average precision of each query and its LOD across MoIR and CLIR on mMARCO.

6 Further Analysis

In Section 6.1, we first establish to what degree the reranking performance is correlated to the lexical overlap between queries and documents. We then investigate the reranking results at the instance level by inspecting individual pair-wise classification results (Section 6.2). Here, we evaluate whether removing lexical overlap causes models to recover from incorrect predictions.

6.1 Correlation Between Lexical Overlaps and Reranking Performance

Table 4 summarizes the ALOD and AP–LOD correlation across MoIR and CLIR on mMARCO. As expected, relevant documents exhibit higher lexical overlap with the query, and this signal is stronger in MoIR (0.90) than CLIR (0.20). Across all models, the AP–LOD correlation is consistently higher in MoIR than CLIR, confirming that MoIR reranking relies more heavily on surface-level overlap. In CLIR, due to vocabulary mismatch between query and document languages, lexical overlap is weak and often limited to named entities, forcing models to rely more on semantic relatedness features.

Among all models, EN-EN tuning shows the strongest correlation between AP and lexical overlap, which means it relies heavily on surface word matching. Instruction-tuning on code-switched data also increases this reliance, though to a lesser extent, suggesting more semantic-driven decisions.

6.2 Causal Effect of Removing Lexical Overlap

Figure 3 and Table 5 summarize the results of models that have been instruction-tuned on the mMARCO dataset. For examples with lexical–semantic conflicts, the `EN-EN-tuned` model shows a recovery rate of 0.500, i.e., half of its false-positive predictions were corrected once lexical overlap cues were removed. This suggests a causal dependence on surface-level keyword overlap. By contrast, the two `Code-switched-tuned` models (`EN-XX-tuned` and `XX-XX-tuned`) show smaller recovery rates (0.300 and 0.455), suggesting that training rerankers on code-switched data indeed mitigates their lexical bias. However, it is important to interpret the results with caution, as the total number of false positives is relatively small.

For the examples without lexical–semantic conflicts, the `Zero-shot` achieves perfect retention (1.00), whereas the `EN-EN-tuned` and `Code-switched-tuned` models show slightly lower scores (0.975–0.995). This indicates that instruction-tuned models still exhibit a slight tendency to rely on lexical overlaps when correctly identifying the relevant passage. This observation aligns with our AP–LOD correlation analysis, where instruction-tuned models show stronger positive correlations between lexical overlap and relevance scores. Different from our reranking results (Section 5), we find that the `Zero-shot` model outperforms instruction-tuned models. This may be explained by domain differences: Both the

Model	Retention Rate Analysis			Recovery Rate Analysis		
	Accuracy	True Positives	Retention Rate	Accuracy	False Positives	Recovery Rate
`Zero-shot`	1.000	204 / 204	1.000	1.000	0 / 200	–
`EN-EN-tuned`	1.000	204 / 204	0.976	0.690	62 / 200	0.500
`EN-XX-tuned`	1.000	204 / 204	0.995	0.900	20 / 200	0.300
`XX-XX-tuned`	0.995	203 / 204	0.995	0.945	11 / 200	0.455

Table 5: Summary of causal analysis. **Left:** Results in terms of classification accuracy, number of instances where models correctly prefer relevant document (true positives; TP), and the fraction of TP instances where models still identify relevant passage after removing lexical overlap (retention rate). **Right:** Results in terms of classification accuracy, number of instances where models incorrectly prefer non-relevant document with lexical overlap (False Positives; FP), and the fraction of FP instances where the preferred passage changes after removing lexical overlap.

`EN-EN-tuned` and `Code-switched-tuned` models were fine-tuned on the mMARCO and XQuAD-R datasets, which improved their in-domain performance but reduced robustness when evaluated on our synthetic data.

Overall, our findings provide causal evidence that lexical overlap directly influences relevance judgments. Compared to the `EN-EN-tuned` model, instruction-tuning on code-switched data reduces but does not fully removes lexical bias.

7 Conclusion

In this study, we investigate to what extent LLM-based rerankers suffer from lexical biases as opposed to semantic relevance. Our results on MoIR and CLIR show that instruction-tuning on English data is most effective for monolingual retrieval, whereas code switching provides the largest benefits in CLIR. We also show that the correlation between reranking performance and lexical overlap is stronger for models trained on monolingual data compared to those trained on code-switched data. Our causal analysis reveals that spurious lexical cues can mislead the model, but their removal often restores correct semantic judgments. These findings highlight both the promise of code-switched data for improving cross-lingual generalization and the need to address lexical bias to ensure that LLMs are "relevant for the right reasons."

8 Limitation and Future Work

Due to the high computational costs of instruction-tuning LLMs, we limit our study to the widely-used Llama-3.1-8B-Instruct model. In addition, the multilingual code-switched data was generated with a fixed translation probability of 0.5, leaving open how different translation probability might affect cross-lingual generalization and lexical bias. In future work, we plan to (1) detect lexical biases at the model-internal level, in order to better understand how lexical overlap reliance and cross-lingual alignment are shaped by different training data, and (2) investigate methods for steering models away from undesired shortcut behavior. Finally, our causal analysis is limited to monolingual examples. We plan to extend this framework to cross-lingual settings in future work.

Acknowledgments

We acknowledge the support for BP through the ERC Consolidator Grant DIALECT 101043235.

References

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Andreas Chari, Iadh Ounis, and Sean MacAvaney. 2025. Lost in transliteration: Bridging the script gap in neural ir. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2900–2905, New York, NY, USA. Association for Computing Machinery.

Yiqun Chen, Qi Liu, Yi Zhang, Weiwei Sun, Xinyu Ma, Wei Yang, Daiting Shi, Jiaxin Mao, and Dawei Yin. 2025. Tourrank: Utilizing large language models for documents ranking with a tournament-inspired strategy. In *THE WEB CONFERENCE 2025*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Changjiang Gao, Hankun Lin, Shujian Huang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Jiajun Chen. 2025. Understanding llms' cross-lingual context retrieval: How good it is and where it comes from. *Preprint*, arXiv:2504.10906.

Lovisa Hagstrom, Ercong Nie, Ruben Halifa, Helmut Schmid, Richard Johansson, and Alexander Junge. 2025. Language model re-rankers are fooled by lexical similarities. *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*.

Donna Harman. 1992. Evaluation issues in information retrieval. *Inf. Process. Manag.*, 28(4):439–440.

Zhiqi Huang, Puxuan Yu, Shauli Ravfogel, and James Allan. 2024. Language concept erasure for language-invariant dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13261–13273, Miami, Florida, USA. Association for Computational Linguistics.

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.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *Preprint*, arXiv:2304.05613.

Napat Laosaengpha, Thanit Tativannarat, Attapol Rutherford, and Ekapol Chuangsuwanich. 2025. Mitigating language bias in cross-lingual job retrieval: A recruitment platform perspective. *Preprint*, arXiv:2502.03220.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Robert Litschko, Ekaterina Artemova, and Barbara Plank. 2023a. Boosting zero-shot cross-lingual retrieval by training on artificially code-switched data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3096–3108, Toronto, Canada. Association for Computational Linguistics.

Robert Litschko, Max Müller-Eberstein, Rob van der Goot, Leon Weber-Genzel, and Barbara Plank. 2023b. Establishing trustworthiness: Rethinking tasks and model evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2022. Streamlining evaluation with ir-measures. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, page 305–310, Berlin, Heidelberg. Springer-Verlag.

Andrianos Michail, Simon Clematide, and Rico Sennrich. 2025. Examining multilingual embedding models cross-lingually through llm-generated adversarial examples. *Preprint*, arXiv:2502.08638.

OpenAI. 2025. GPT-5 System Card. Technical Report Technical Report, OpenAI. Technical report; accessed: 06 October 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Andrew Parry, Sean MacAvaney, and Debasis Ganguly. 2024. Top-down partitioning for efficient list-wise ranking. *Preprint*, arXiv:2405.14589.

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.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, 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.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6883–6893, Torino, Italia. ELRA and ICCL.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2327–2340, Mexico City, Mexico. Association for Computational Linguistics.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Comput. Surv.*, 56(12).

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

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

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.

Jerrold H. Zar. 2005. *Spearman Rank Correlation*. John Wiley & Sons, Ltd.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source large language models are strong zero-shot query likelihood models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8807–8817, Singapore. Association for Computational Linguistics.

Longfei Zuo, Pingjun Hong, Oliver Kraus, Barbara Plank, and Robert Litschko. 2025. Evaluating large language models for cross-lingual retrieval. *Preprint*, arXiv:2509.14749.

A Pairwise Re-ranking Prompt Template

System Prompt

```
You are an expert in multilingual
information retrieval. Your task is to
determine which of the two passages is more
relevant to the given query.
Strict instructions:
- Do NOT provide any explanation.
- Do NOT include any additional words,
punctuation, or formatting.
- Answer with only Passage A or Passage B
(without quotes).
```

User Prompt

```
Query: {query}
Passage A: {doc1}
Passage B: {doc2}
Which passage is more relevant to the query?
Respond with exactly one of the following
options:
Passage A
Passage B
Your answer:
```

Figure 4: Prompt for pairwise re-ranking.

B Hyperparameters and Infrastructure

Hyperparameter	Value
Maximum sequence length	1024
Learning rate	2e-5
Batch size	32
Warm-up ratio	0.03
Optimizer	AdamW (Loshchilov and Hutter, 2017)
Re-ranking Model	`Llama-3.1-8B-Instruct`
LLM Parameters	8 Billion

Table 6: Hyperparameter values for re-ranking models used in our experiments.

Setup	Value
GPU	NVIDIA H100 SXM5-GPUs (94 GB)
Avg. Training Duration (per model)	45 h
Avg. Test Duration (per language pair)	87 h

Table 7: Computational environment. We use the Huggingface framework to train our models (von Werra et al., 2020), `ir-measures` for computing MRR@10 (MacAvaney et al., 2022), and Spearman correlation coefficients for correlation analysis.

C Prompting Format

```
<|begin_of_text|><|start_header_id|>system<|
end_header_id|>
You are an expert in multilingual
information retrieval. Your task is to
determine which of the two passages is more
relevant to the given query.
<|eot_id|><|start_header_id|>user<|
end_header_id|>
Query: ....
Passage A: ....
Passage B: ....
Which passage is more relevant to the query?
Your answer:<|eot_id|><|start_header_id|>
assistant<|end_header_id|>
```

Figure 5: A simplified example of a chat-formatted prompt using the official LLaMA-3.1 chat template. This example is only for illustration and does not reflect the full prompt used in our experiments. For the complete prompt we use, see Appendix A.

D GPT-5 Synthetic Data Generation Prompts

This appendix provides the exact GPT-5 prompt templates used for generating and perturbing the synthetic data described in Section 3.3. All prompts are shown in their natural-language form for reproducibility.

Prompt 1: Lexical-Semantic Non-Conflict Candidate Generation

```
Please generate 20 samples in jsonl
format for pairwise semantic relevance
reranking task. Each sample must follow
the content requirements and format
requirements below.

**Content requirements:**
(1) The query should be a "wh"-question
and keywords in the questions must have
synonyms.
(2) Passage A must always be semantically
 relevant to the query. Passage B must
always be semantically irrelevant to the
query.
(3) Passage A and the query must share at
 least one overlapping non-stopword
keyword. Passage B must not contain any
overlapping token with the query.
(4) Passage A and Passage B should be
about similar or related topics, so that
the pair forms a hard example (difficult
to judge at first glance, but with a
unique correct answer).

**Format requirements:**
(1) Output must be in jsonl format.
(2) Each entry must include: qid, query,
passage_A, passage_B, and output.
(3) Each qid and pid must be unique and
assigned in order.
(4) Always set "output": "Passage A".
```

```
Now, please directly generate 20 new
samples that strictly follow the above
rules.
```

Figure 6: GPT-5 prompt used for generating lexical–semantic **non-conflict (TP)** examples.

Prompt 2: Lexical-Semantic Conflict Candidate Generation

```
Please generate 20 samples in jsonl
format for pairwise semantic relevance
reranking task. Each sample must follow
the content requirements and format
requirements below.

**Content requirements:**
(1) The query should be a "wh"-question
and keywords in the questions must have
synonyms.
(2) Passage A must always be semantically
 relevant to the query. Passage B must
always be semantically irrelevant to the
query.
(3) Passage B and the query must share at
 least one overlapping non-stopword token.
 **Passage A** must not contain any
overlapping token with the query.
(4) Passage A and Passage B should be
about similar or related topics, so that
the pair forms a hard example (difficult
to judge at first glance, but with a
unique correct answer).

**Format requirements:**
(1)Output must be in JSONL format.
(2)Each entry must include: qid, query,
passage_A, passage_B, and output.
(3)Each qid and pid must be unique and
assigned in order.
(4)Always set "output": "Passage A".

Now, please directly generate 20 new
jsonl samples that strictly follow the
above rules.
```

Figure 7: GPT-5 prompt used for generating lexical–semantic **conflict** (FP) examples.

Prompt 3: Lexical-semantic Non-Conflict True Positive Example Perturbation

```
Please perturb each of the following
triples (original examples) used for
pairwise semantic relevance reranking.
These examples all satisfy the following
conditions:
(1) "gold_output" Passage is always
semantically relevant to the query. the
other passage is always semantically
irrelevant to the query.
(2) the **relevant passage** and the
query share at least one overlapping non-
stopword token.

**Perturbation requirements:**
(1) Replace ALL OVERLAPPING TOKENS in the
 **Relevant Passage** that also appears
in the query (i.e., all overlapping
tokens) with suitable synonyms, while
keeping the overall sentence semantics
unchanged.
(2) Do not modify any other part of
relevant passage except the overlapping
tokens, and make sure all overlapping
tokens are replaced. Do not modify
irrelevant passage.
(3) The output must be in JSONL format,
consistent with the structure of the
original examples.

Following the above instructions, please
perturb those original examples provided
below and return the results in JSONL
format.
```

Figure 8: GPT-5 prompt used for perturbing lexical–semantic **non-conflict (TP)** examples.

Prompt 4: Lexical-semantic Conflic False Positive Example Perturbation

```
Please perturb each of the following
triples (original examples) used for
pairwise semantic relevance reranking.
These examples all satisfy the following
conditions:
(1) "gold_output" Passage is always
semantically relevant to the query. the
other passage is always semantically
irrelevant to the query.
(2) the **irrelevant passage** and the
query share at least one overlapping non-
stopword token.

**Perturbation requirements:**
(1) Replace all overlapping tokens in **
irrelevant passage** that also appears in
 the query (i.e., all overlapping tokens)
 with suitable synonyms, while keeping
the overall sentence semantics unchanged.
(2) Do not modify any other part of
irrelevant passage except the overlapping
 tokens, and make sure all overlapping
tokens are replaced. Do not modify
relevant passage.
(3) The output must be in JSONL format,
consistent with the structure of the
original examples.

Following the above instructions, please
perturb the 20 original examples provided
 below and return the results in JSONL
format.
```

Figure 9: GPT-5 prompt used for perturbing lexical–semantic **conflict (FP)** examples.

	EN-EN			EN-DE			EN-IT			EN-AR		
	0	[1, 3)	[3, +∞)	0	[1, 3)	[3, +∞)	0	[1, 3)	[3, +∞)	0	[1, 3)	[3, +∞)
Zero-shot	83.1	87.9	**83.5**	88.4	87.9	**85.1**	88.2	**88.1**	90.9	87.2	**86.5**	**86.5**
EN-EN-tuned	95.3	95.1	**90.3**	91.2	91.1	**79.3**	92.1	89.9	**82.6**	87.5	86.2	**73.0**
EN-XX-tuned	96.6	97.4	**92.2**	93.6	94.0	**86.2**	93.8	93.3	**87.6**	89.9	91.0	**82.4**
XX-XX-tuned	94.6	96.5	**94.2**	94.0	94.0	**93.1**	94.5	93.8	**90.9**	90.7	91.1	**83.8**

Table 8: Accuracy of pairwise relevance classification on the mMARCO dataset, where models are prompted to judge which of two passages is more relevant to a query. The relevant passage is lexically disjoint from the query, while the irrelevant passage exhibits varying degrees of lexical overlap. Irrelevant passages are grouped into three categories based on their overlap count with the query: 0 (no overlap), [1, 2) (low overlap), and [3, +∞) (high overlap). The table reports classification accuracy across language pairs and overlap levels. **Bold** indicates the overlap group with the lowest accuracy for each model–language-pair pair.

	EN-EN	EN-DE	EN-IT	EN-AR
Zero-shot	32.4	23.1	36.4	70.0
EN-EN-tuned	55.0	50.0	38.1	35.0
EN-XX-tuned	50.0	33.3	46.7	46.2
XX-XX-tuned	50.0	16.7	36.4	25.0

Table 9: Accuracy (recovery rate) of different models on the subset of triple samples where the irrelevant document originally had ≥3 lexical overlaps with the query and was incorrectly predicted as relevant. Results shows the proportion of cases in which models correctly identified the relevant document after removing the overlapping tokens.

E Causal Analysis with Word2Vec-based Perturbation

Overlap	EN-EN	EN-DE	EN-IT	EN-AR
0	296	32,157	31,457	53,664
[1, 3)	1,759	7,218	6,262	4,652
[3, +∞)	206	87	121	74

Table 10: Number of pair-wise classification instances extracted from mMARCO, grouped by how many tokens overlap between the query and non-relevant document.

In an earlier version of our causal analysis, we used real examples from the mMARCO dataset and applied word2vec-based perturbations. we first identify triplets $u(q, d_r, d_{\text{nr}})$ where the relevant document d_{r} shares no overlap with the query, while the non-relevant document d_{nr} contains varying degrees of overlap. Inspired by (Litschko et al., 2023a), we partition samples into those where d_{nr} has no overlap (0 tokens), low overlap (1–2 tokens), and high overlap (≥3 tokens) with q (see Table 10). For these high-overlap samples, we replaced overlapping tokens in the non-relevant document with their nearest neighbors in the word2vec embedding space and re-evaluated model predictions.

Table 8 shows the classification accuracy of all four models across the four language pairs (dubbed *clean run*). Among the 16 combinations of 4 language pairs and 4 models, we observed a consistent pattern: in 12 of these settings, classification accuracy is lowest when the number of overlapping tokens between the query and non-relevant document was greater than or equal to three. For example, for the XX-XX-tuned model on the EN-AR pair, the accuracy falls to 0.838 in the high-overlap group, while reaching 0.907 in the no-overlap group and 0.911 in the low-overlap group. The only exceptions were the zero-shot model applied to the EN-EN and EN-IT language pairs. These drops suggest that models are more likely to over-rely on lexical overlap signals, leading to misclassification when the overlap is misleading.

When comparing models within the same language pair, we find that in cases where the irrelevant document shares at least one token with the query, the models trained on code-switched data generally outperformed the EN-EN-tuned model. For example, for the CLIR pair EN-IT, the EN-XX-tuned and XX-XX-tuned models reach accuracies of 0.876 and 0.909, respectively, exceeding the EN-EN-tuned model's performance of 0.826.

In the second part of the experiment, we focus on misclassified samples from each language pair in the $[3, \infty)$ group and measure if substituting overlapping tokens causes the predictions to change.

	stopword removal	lemmatization	subword tokenizer	without negatives	top-5 negatives	top-10 negatives	top-20 negatives	top-50 negatives
MoIR ALOD	0.847	1.083	2.042	2.508	0.902	0.903	0.904	0.904
CLIR ALOD	0.242	0.235	0.460	0.406	0.197	0.197	0.196	0.197

Table 11: Robustness analysis of the ALOD metric under different preprocessing alternatives and varying number of negative documents extracted top the top-k documents in the input ranking.

Table 9 quantifies this recovery effect by reporting the proportion of misclassified high-overlap samples that were corrected after corruption. We observe that all four models show improved accuracy on these modified samples across all language pairs.

However, upon closer inspection, we found that this setup had several limitations. Many overlapping tokens corresponded to named entities or fixed expressions whose substitution could not preserve meaning, and word2vec neighbors sometimes introduced semantic drift. To ensure more controlled perturbations and consistent semantics, we therefore replaced this analysis with the synthetic GPT-5–generated data described in Section 3.3, which allows for precise manipulation of lexical overlap while maintaining contextual coherence.

F ALOD Robustness: Experimental Results

This appendix reports the robustness evaluation results for the ALOD metric, as summarized in Table 11.