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

Yuchen Mao[1*], Barbara Plank[23], Robert Litschko[23]

yuchmao@lst.uni-saarland.de, {b.plank, robert.litschko}@lmu.de

[1] Department of Language Science and Technology, Saarland University, Germany
[2] Center for Information and Language Processing (CIS), LMU Munich, Germany
[3] Munich Center for Machine Learning (MCML), Munich, Germany

Introduction & Motivation

- **Problem:**

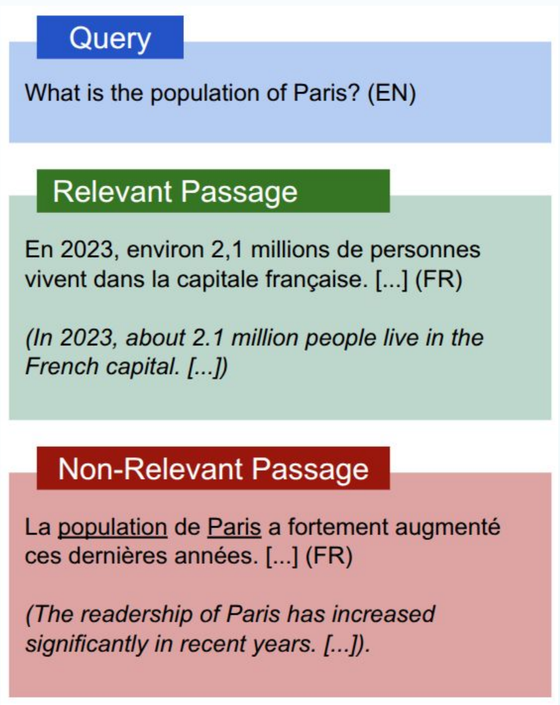
Large Language Models (LLMs) are powerful rerankers in Information Retrieval (IR), but training only on monolingual data often causes **monolingual overfitting** and **lexical bias**, limiting cross-lingual generalization.

- **Key Research Question:**

Are LLM rerankers relevant for the right reasons (or are they just matching words rather than meaning)?

- **Example:**

LLMs may prefer lexically overlapping but semantically irrelevant passages. 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.



Methodology & Experiments

Model & Baselines:

- Base Model: LLaMA-3.1-8B-Instruct
- Variants:
  - Zero-shot (without fine-tuning)
  - EN–EN tuned (monolingual English)
  - Fine-tuned (target language pairs)
  - EN–XX tuned (code-switched queries)
  - XX–XX tuned (code-switched queries + docs)

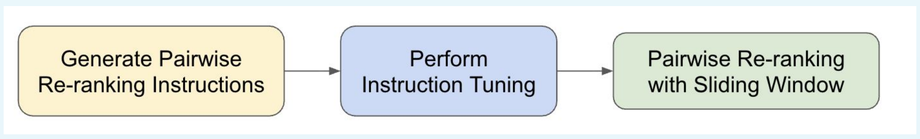
The **clima en** Tokyo **es muy** warm **durante verano**.
(The climate in Tokyo is very warm during summer.)

An example of multilingual code-switch.

- **Data:** multilingual MSMARCO (mMARCO), XQuAD-R

(1) Instruction-tuning and Reranking Pipeline

- **Tasks:** MoIR (Monolingual IR), CLIR (Cross-Lingual IR)



- **Measuring Lexical Bias**
  - ALOD: $LOD_q = \frac{1}{|D_q^+|} \sum_{d \in D_q^+} \text{Overlap}(q,d) - \frac{1}{|D_q^-|} \sum_{d \in D_q^-} \text{Overlap}(q,d)$

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

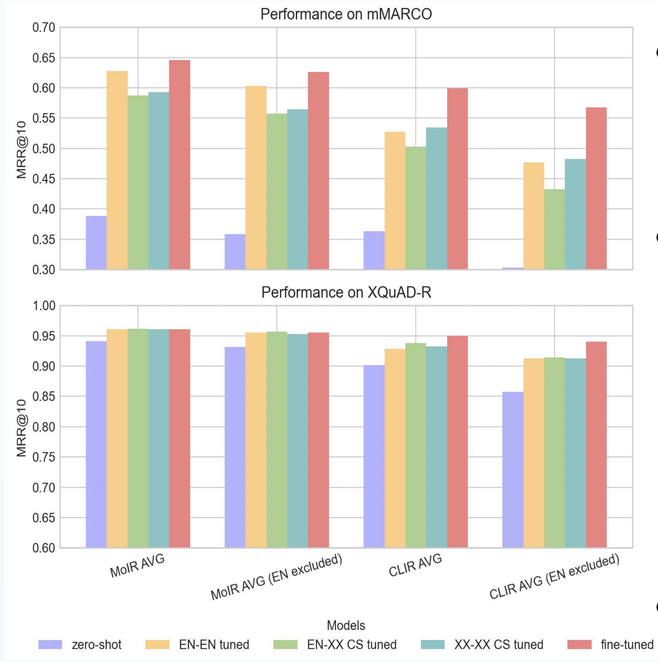
  - AP-LOD correlation: the Spearman correlation between the average precision of each query and its LOD.
  - **High ALOD → higher potential lexical bias**
  - **High correlation → model performance depends on overlap**

(2) Causal Analysis

- **Retention Rate** (no lexical-semantic conflict):
  - Measures how often models stay correct after removing lexical overlap
  - → higher = less lexical bias.
- **Recovery Rate** (lexical-semantic conflict):
  - Measures how often models correct wrong predictions after overlap removal
  - → higher = more lexical bias.

Results Analysis

- **Reranking Results**



- MoIR vs. CLIR:
  - Re-ranking performance is consistently higher in monolingual settings across all models.
- Training strategy comparison:
  - The EN–EN tuned model performs better on MoIR (on mMARCO), while the ML–CS tuned model performs better on CLIR.
  - Fine-tuned models remain best overall, zero-shot the weakest.
- English-centric bias persists.

- **Lexical Bias Analyses**

	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

- Relevant documents exhibit higher lexical overlap with the query, and this signal is stronger in MoIR (0.90) than CLIR (0.20).
- AP-LOD correlation is significantly stronger in MoIR than CLIR → MoIR re-ranking relies more heavily on surface-level overlap.

- **Causal Analysis**

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

- Recovery Rate: After the lexical overlap removal and synonym replacement, all three models show improved accuracy on these modified FP samples, and code-switched models less biased.
- Retention Rate: Instruction-tuned models show mild reliance on lexical overlap.

Conclusion

- **Effect of Instruction Tuning:**
  - English-only tuning leads to monolingual overfitting: models perform best on English (MoIR) but struggle to generalize across languages.
  - Code-switched tuning improves cross-lingual robustness but slightly reduces monolingual precision compared to English-only tuning.
  - Instruction tuning boosts in-domain accuracy, yet reinforces reliance on surface lexical cues.
- **Lexical vs Semantic Relevance:**
  - Removing shared words reveals causal dependence on lexical overlap: models often match words rather than meaning.
  - Code-switched models rely less on surface cues, showing better semantic generalization.
  - ⇒ Models are not always "right for the right reasons."

Acknowledgements

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

* Work done while at LMU Munich.