Evaluating Large Language Models for Cross-Lingual Retrieval

Longfei Zuo*[1] Pingjun Hong*[1] Oliver Kraus[1] Barbara Plank[1,2] Robert Litschko[1,2]

[1] MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

[2] Munich Center for Machine Learning (MCML), Munich, Germany

{zuo.longfei, pingjun.hong}@campus.lmu.de, {o.kraus2, b.plank, robert.litschko}@lmu.de

EMNLP 2025
Suzhou, China | 中国苏州

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Munich Center for Machine Learning

Overview

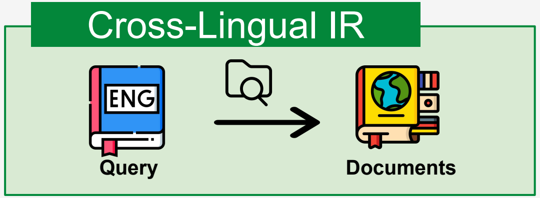
- Introduction
- Research Questions
- Experimental Setup
- Results
- Conclusions

Overview

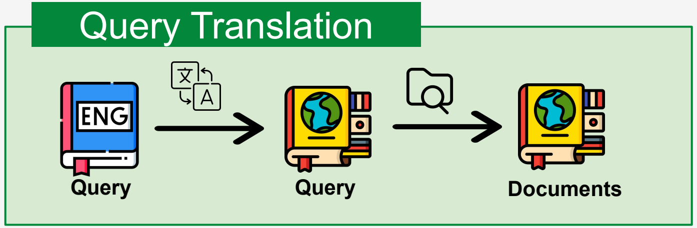
- **Introduction**
- Research Questions
- Experimental Setup
- Results
- Conclusions

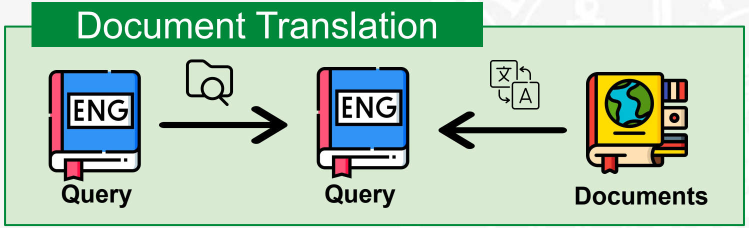
Introduction

- 🌏 **Cross-Lingual IR**→ retrieve documents in a different language than the query



- Previous evaluation setup relies on lexical retrieval with machine translation (MT).

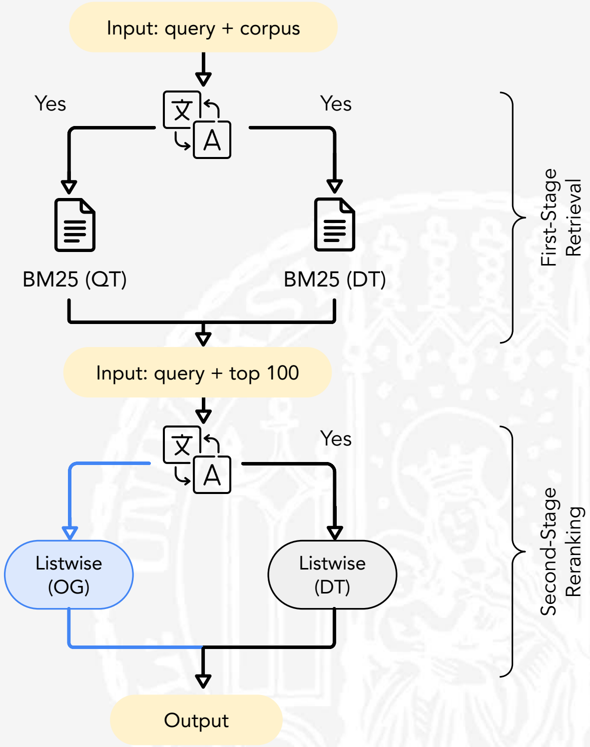




- **Challenges:** *Latency*, *Low-resource issues, Translation errors*

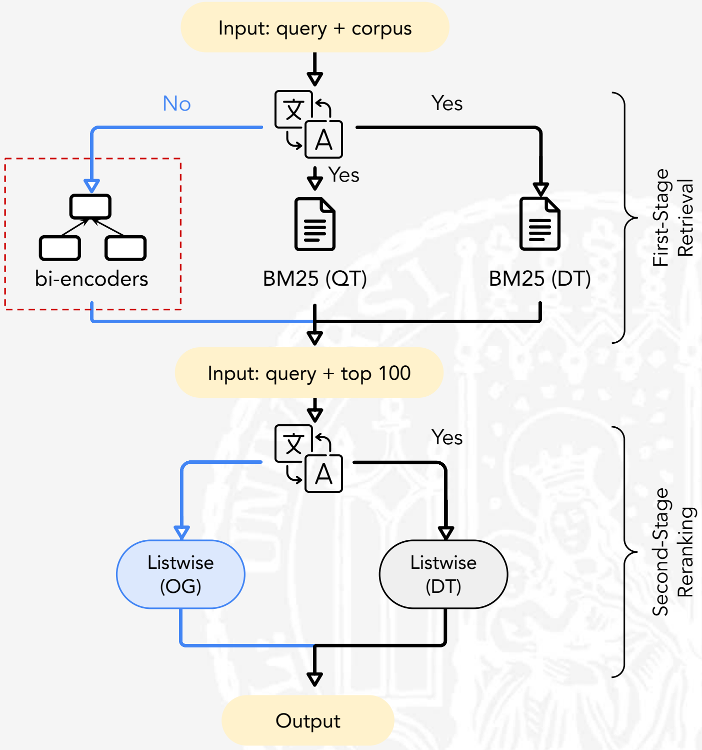
Introduction

- **Most similar work** (Adeyemi et al., 2024)
  - ➔ *First-stage: Query or document translation with BM25*
  - ➔ *Reranking: Listwise rerankers*
  - ➔ *Focus on African languages (CIRAL)*



Introduction

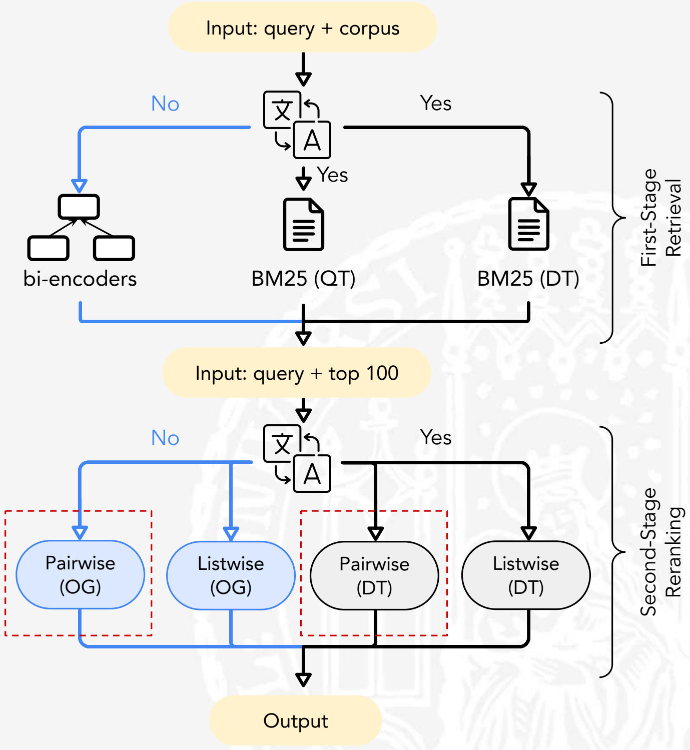
- **Most similar work** [(Adeyemi et al., 2024)]
  - ➔ *First-stage: Query or document translation with BM25*
  - ➔ *Reranking: Listwise rerankers*
  - ➔ *Focus on African languages (CIRAL)*
- **We extend this study:**
  - ➔ *First-stage: Multilingual bi-encoders + BM25*



Icons by Freepik from www.flaticon.com.

Introduction

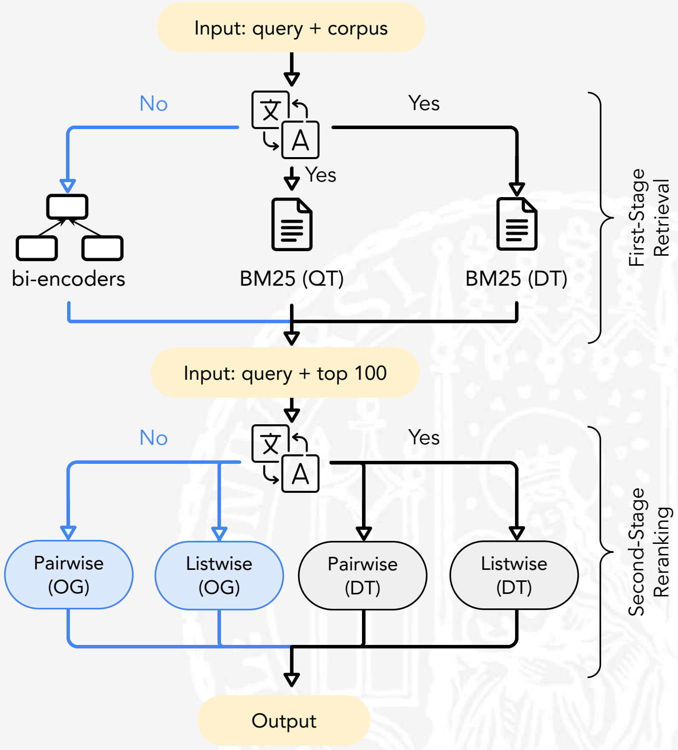
- **Most similar work** [(Adeyemi et al., 2024)]
  - ➔ *First-stage: Query or document translation with BM25*
  - ➔ *Reranking: Listwise rerankers*
  - ➔ *Focus on African languages (CIRAL)*
- **We extend this study:**
  - ➔ *First-stage: Multilingual bi-encoders + BM25*
  - ➔ *Reranking: Listwise and pairwise rerankers*



Icons by Freepik from www.flaticon.com.

Introduction

- **Most similar work** [(Adeyemi et al., 2024)]
  - ➔ *First-stage: Query or document translation with BM25*
  - ➔ *Reranking: Listwise rerankers*
  - ➔ *Focus on African languages (CIRAL)*
- **We extend this study:**
  - ➔ *First-stage: Multilingual bi-encoders + BM25*
  - ➔ *Reranking: Listwise and pairwise rerankers*
  - ➔ *African and European languages (CLEF 2003)*



Icons by Freepik from www.flaticon.com.

Overview

- Introduction
- **Research Questions**
- Experimental Setup
- Results
- Conclusions

Research Questions

- **RQ1:** How do recent multilingual bi-encoders and translation-based sparse retrieval compare in their performance?
- **RQ2:** Do LLM-based rerankers improve CLIR performance across high-resource and low-resource languages?
- **RQ3:** How do pairwise and listwise rerankers influence CLIR performance?
- **RQ4:** What is the impact of document length on reranking in CLIR?

Overview

- Introduction
- Research Questions
- **Experimental Setup**
- Results
- Conclusions

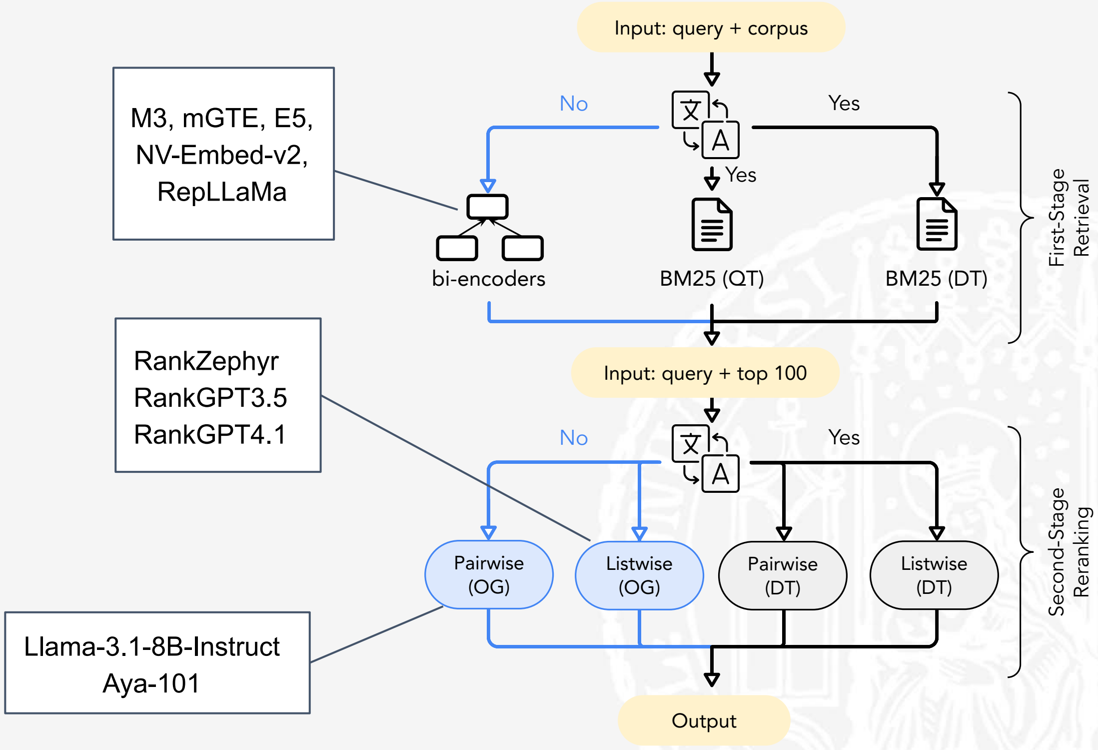
Experimental Setup

Datasets

- CLEF 2003 (Braschler, 2003)
  - EN–{DE, IT, FI, RU}
  - DE–{FI, IT, RU}
  - FI–{IT, RU}
- CIRAL (Adeyemi, 2024)
  - EN–{HA, SO, SW, YO}

Document translation

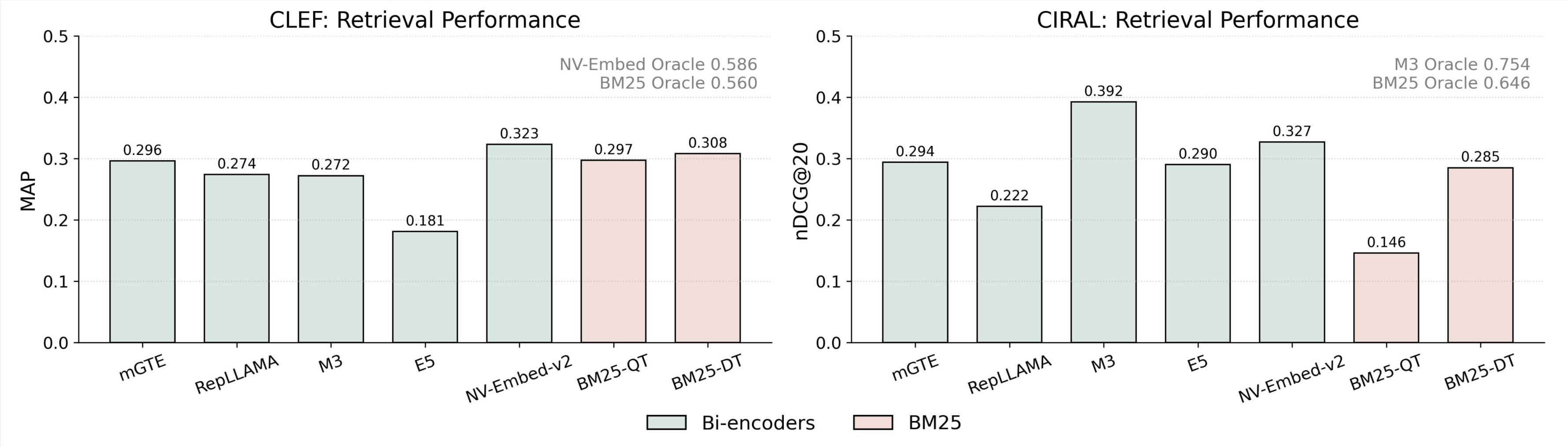
- Sentence-by-sentence translation with NLLB (NLLB Team, 2022)
- Model: nllb-200-1.3B



Overview

- Introduction
- Research Questions
- Experimental Setup
- **Results**
- Conclusions

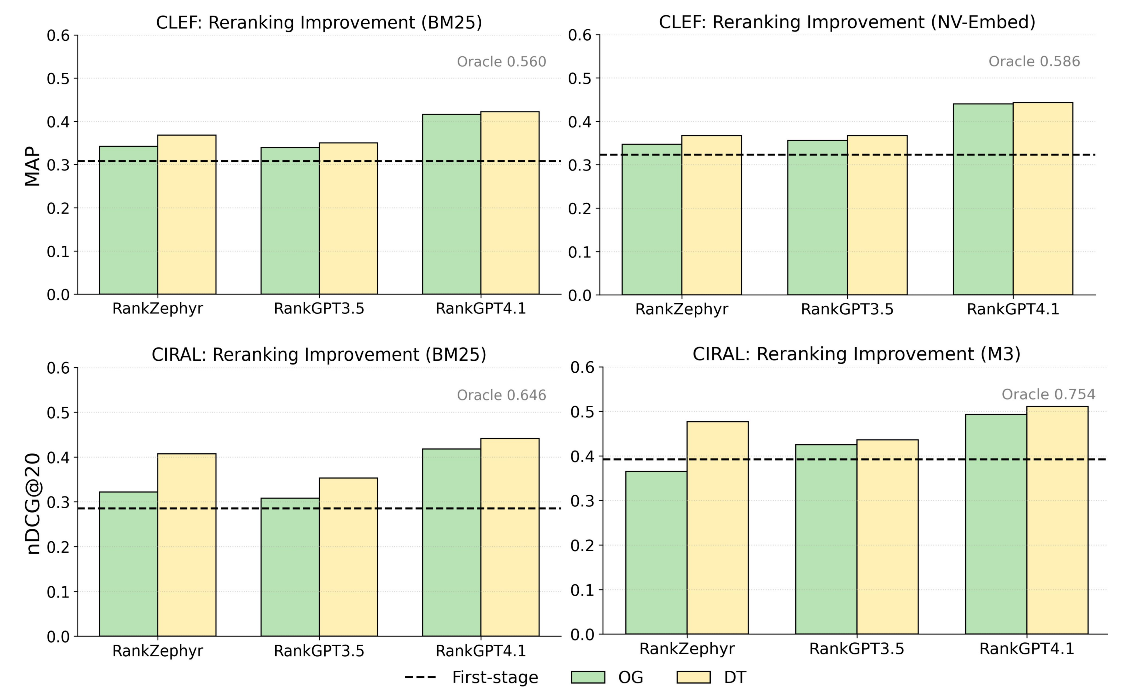
First-stage Retrieval



Retrieval Results (RQ1)

- The best-performing retriever differs considerably between the datasets.
- Recent bi-encoders achieve superior performance.

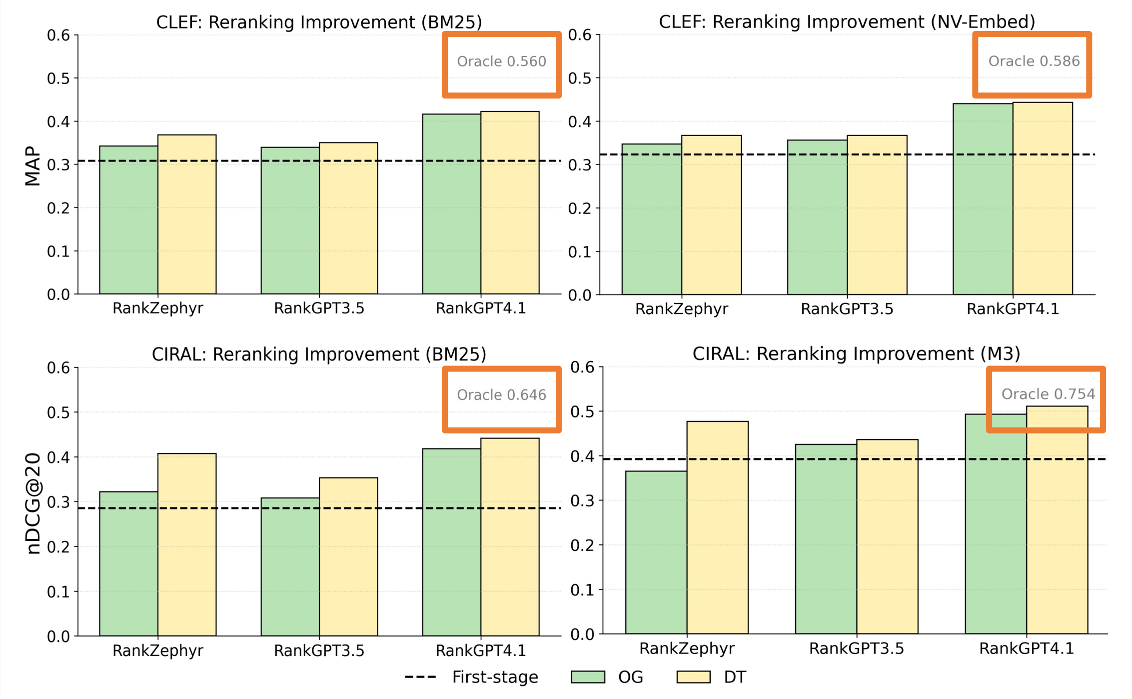
Second-Stage Listwise Reranking



Reranking Results (RQ2)

- Consistent across both datasets: better retrieval yields better reranking.
- The benefits from DT appears to diminish, as models achieve stronger performance.

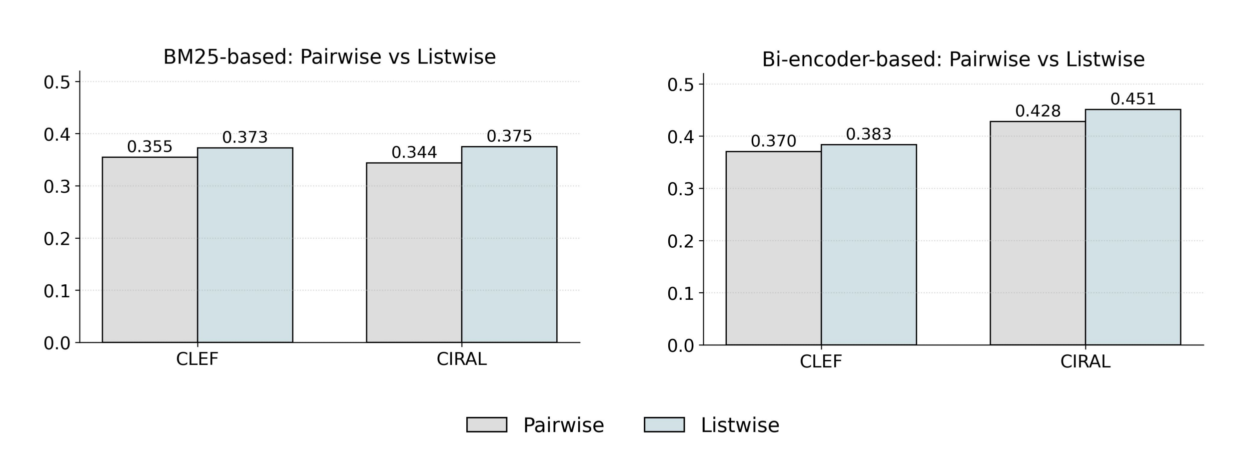
Second-Stage Listwise Reranking



The Glass Ceiling of Reranking

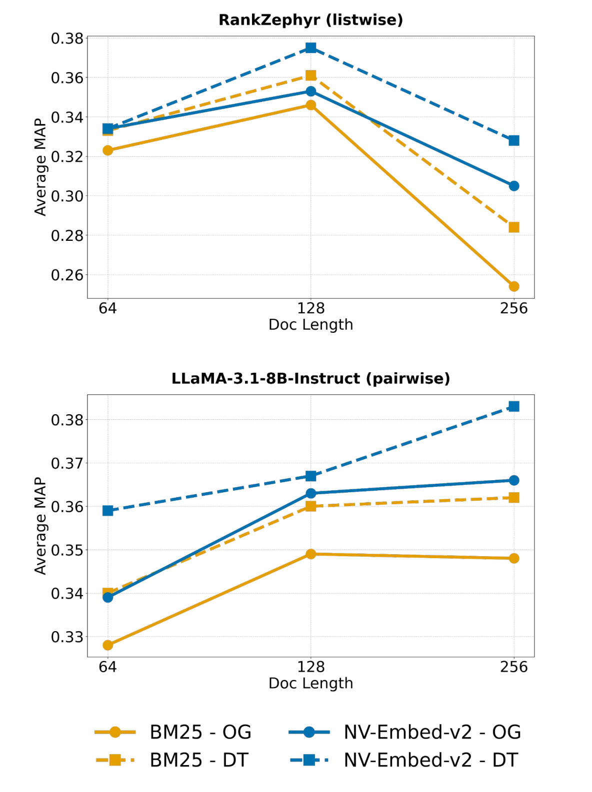
- **Oracle:** best possible result based on input ranking.
- **"Ceiling Effect":** State-of-the-art rerankers struggle to approach the oracle.

Listwise vs. Pairwise Reranking (RQ3)



- Instruction-tuned pairwise rerankers perform competitively with listwise rerankers.

Impact of document length



- RankZephyr performs best at 128 tokens.
  - Listwise rerankers are sensitive to overly long inputs.
- Using BM25, the gains from longer input saturate or even slightly decline.
  - Pairwise reranker, when paired with high-quality dense retriever, is more robust to longer input spans.

Optimal input length is task- and model-dependent and may require careful tuning (RQ4).

Overview

- Introduction
- Research Questions
- Experimental Setup
- Results
- **Conclusions**

Conclusion

- Improved retrieval results translates to better reranking results, the benefits of MT diminish with stronger listwise rerankers.

- Pairwise rerankers, based on instruction-tuned LLMs, perform competitive to listwise rerankers trained on retrieval supervision.

- Performance varies with data conditions — reranking performance is sensitive to document lengths and language pairs.

Thank you for your attention!

Paper **Code**