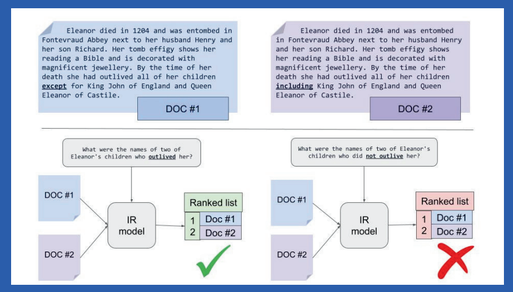
Reproducing NevIR: Negation in Neural Information Retrieval

Negation is a fundamental aspect of human communication, yet it remains a challenge for Language Models (LMs) in Information Retrieval (IR). Despite the heavy reliance of modern neural IR systems on LMs, little attention has been given to their handling of negation.

Coen van den Elsen, Francien Barkhof, Thijmen Nijdam
Supervised by: Simon Lupart, Mohammad Alianne Jadi

SIGIR 2025 Padova

NevIR Benchmark



The NevIR dataset consists of contrastive query-document pairs with identical documents except for a key negation. Each query aligns semantically with one of the two documents, where Query #1 corresponds to Doc #1, and Query #2 to Doc #2.

Evaluation metric: pairwise accuracy

Models

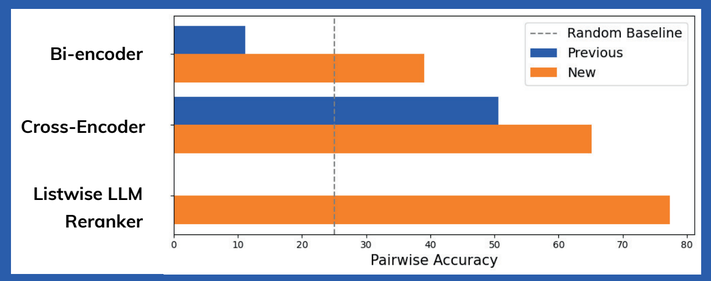
Random baseline: Pairwise accuracy is 25% (½ x ½ = ¼)

Bi-Encoders: Encode queries and documents into single-vector representations

Cross-Encoders: Encode the document and the query together

★ (new) **Listwise LLM re-rankers:** Ask an LLM to reorder a list of retrieved documents

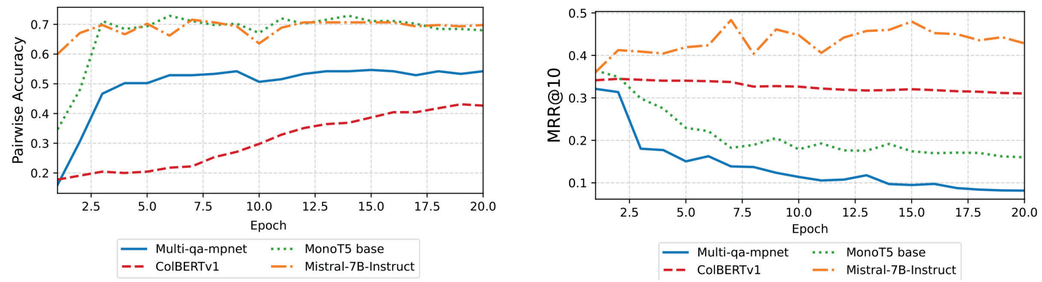
Current IR model negation understanding?



- Across all categories, a SOTA model outperforms previous ones on the NevIR benchmark, with the biggest improvement seen in the bi-encoder category
- The new category, listwise LLM re-ranker, demonstrates the best performance on the NevIR benchmark
- The newly tested cross-encoder has substantially fewer parameters than the previous model, yet performs significantly better

How does finetuning effect negation understanding?

- Across all model categories, finetuning on NevIR improves performance
- The listwise LLM re-ranker retains its ranking performance, while the cross-encoder and bi-encoder do not



Does finetuning generalise across another negation dataset?

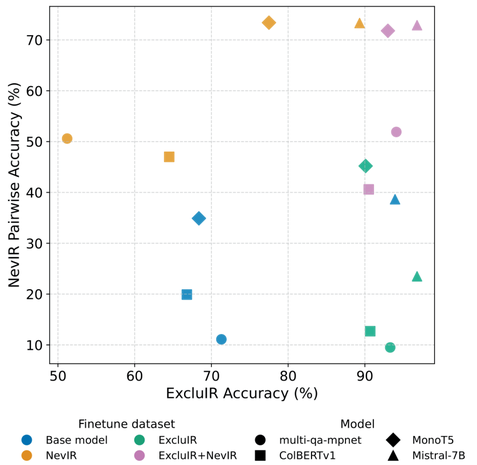
ExcluIR Benchmark

Unlike the NevIR dataset, which primarily focuses on understanding negation semantics within documents, ExcluIR emphasises the exclusionary nature of queries.

"What other flowers **besides** tulips grow in the Netherlands?"

Evaluation metric: Accuracy

- The cross-encoder is the only model that generalises its negation understanding to both datasets
- Finetuning on a merged dataset results in best overall performance



Conclusion

We reproduced and extended NevIR, confirming that most neural IR models still fail to handle negation effectively. Listwise LLM re-rankers have the strongest performance, but at high computational cost. Negation understanding learned on one dataset rarely transfers to another, with only cross-encoders showing robust generalisation. Fine-tuning boosts negation performance but risks overfitting. Overall, negation remains a persistent challenge for neural IR systems.