Aligning Offline Evaluation with Online Performance in Information Retrieval Systems: Query Selection for Alignment

Author
Stephan Popov - skpopov@tudelft.nl

Supervisors
Avishek Anand

01 Background

- Search engines are central to how people access information, and the ranking algorithms behind them play a key role in shaping user experience.
- When new ranking models are proposed, researchers need reliable ways to determine whether these models will improve what users actually see and interact with:
  - 1. **Offline Evlaluation**: testing search quality using pre-collected queries, documents, and human relevance judgments—without involving real users or live traffic.
  - 2. **Online Evaluation:** testing search quality with real users in a live system
- Recent work with large language models has introduced a possible new direction. **Large-Language Models (LLMs)** can generate relevance labels at scale, reducing the need for costly human annotation.[1]

Famous Search Engines in 2026

- At the same time, the choice of queries used for evaluation has become increasingly important. Datasets like MS MARCO contain hundreds of thousands of queries, but evaluation is usually performed on a small subset

02 Aim & Research Questions

Common Aim:

- How do we know if a search engine is giving good results?
- Offline evaluation is cheap and safe, but unreliable:
  - Different selected queries → Different conclusions
- LLM-generated labels add new behaviour and may shift what looks "easy", "hard", or "useful"

Research Sub-Question:

- RQ1: Query selection for alignment. How does the choice of query influence the alignment between offline and online evaluation of ranking systems?

03 Methodology

Dataset
- Rank-DistiLLM[2] (MS MARCO[3]–derived)
  - 1,999 queries
  - 236k relevance labels (CSE3000, LLM-generated)
  - 52 retrieval systems (lexical, dense, sparse, multi-vector, rerankers)
- Oracle relevance labels are used only for comparison to study how label sources affect alignment.

Retrieval Systems
- Experiments use the full pool of 52 systems, then Top-10 systems are selected for computing Kendall's τ
  → reduces noise and improves rank-order stability.

Offline vs. Online Evaluation
- Offline score: computed on a sampled subset of queries Q′
- Simulated online score: computed on the full query set Q
- Alignment metric: Kendall's τ
- Maximise

$$\tau\big(\text{offline}(Q'),\ \text{online}(Q)\big)$$

Purpose of Methodology
- To isolate how query choice affects offline–online alignment and to determine whether different label sources (LLM vs. Oracle) change the behaviour of sampling strategies.

04 Experiments

Query Selection Strategies
1. Random Sampling
→ Baseline for comparison.

2. Query Length
Short / Medium / Long queries (based on word count).
→ Tests whether surface complexity influences alignment.

3. Overlap-Based
Using Jaccard similarity across systems:
- High-overlap: systems agree
- Low-overlap: systems disagree

→ Tests whether system agreement/disagreement is informative.

4. Difficulty-Based
Based on average offline relevance:
- Easy: high relevance
- Hard: low relevance

→ Tests whether "difficulty" affects stability.

Subset Sizes
- For most methods: 5, 10, 25, 50, 100 queries
- For query-length: 5, 10, 15, 20, 25
- Random sampling and query-length repeated 5× (fixed seeds).

05 Results

1. Random Sampling
Kendall's τ varies around 0, but performs better than other rankers when using more retrieval systems (52 vs. 10).
With high system diversity, random subsets capture a broader range of behaviours → becomes a stronger baseline.

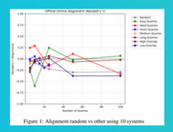
Figure 1: Alignment random vs other using 10 systems

Random sampling vs other algorithms comparison

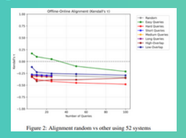
Figure 2: Alignment random vs other using 52 systems

2. Difficulty-Based Sampling (Easy vs. Hard)
Easy queries → consistently positive τ.
Hard queries → τ near 0 (unstable alignment).
When relevance labels are dense (as in CSE3000), systems more often retrieve relevant documents, making easy queries more stable and predictable.

3. Overlap-Based Sampling (High vs. Low Overlap)
High-overlap queries → positive alignment.
With dense labels, system agreement leads to smooth, stable score patterns → higher τ.


Overlap comparison per labels (CSE3000 vs Oracle)

4. Query Length Sampling
Short, medium, and long queries all show τ ≈ −0.1 to +0.1.
No meaningful difference → surface query complexity does not predict alignment.

5. Label Dependency (CSE3000 vs. Oracle Labels)
Oracle labels are sparse → metrics become highly sensitive to disagreement.
→ Low-overlap performs better with Oracle labels because disagreement exposes real differences.
CSE3000 labels are dense → metrics become smoother.
→ High-overlap + easy queries perform better because agreement leads to stable score patterns.

06 Discussion and Conclusion

- Offline–online alignment is highly sensitive to query choice.
- System diversity matters: with many rankers, random sampling becomes a strong baseline.
- Label sources matter:
  - Sparse labels (Oracle) → low-overlap queries work better.
  - Denser LLM labels → high-overlap and easy queries perform better.
- No strategy is universally optimal: performance depends on label distribution, system behaviour, and subset size.

References

- [1] Faggioli, G., Oosterhuis, H., & de Rijke, M. (2023). On the Reliability of LLM-based Relevance Judgments. Proceedings of the 46th International ACM SIGIR Conference.
- [2] Rank-DistiLLM: Schlatt, F., Fröbe, M., Scells, H., Zhuang, S., Koopman, B., Zuccon, G., Stein, B., Potthast, M., & Hagen, M. (2025). Rank-DistiLLM: Closing the Effectiveness Gap Between Cross-Encoders and LLMs for Passage Re-ranking.Proceedings of ECIR 2025.
- [3] MS MARCO: Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., et al. (2016). MS MARCO: A Human-Generated Machine Reading Comprehension Dataset. arXiv:1611.09268.