MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages: 833-836

Empirical Analysis of Score Fusion Strategies under Pool-Restricted Dense Encoding for Ad Hoc Document Retrieval

Syed Minnatullah Quadri[1], Vrishali A. Chakkarwar[1]
[1]Department of Computer Science and Engineering,
Government Engineering College, Aurangabad, Maharashtra 431001, India
E-mail: mt24f05f001@geca.ac.in[1], vachakkarwar@geca.ac.in[2]

Abstract

This paper examines whether restricting dense encoding to the BM25 candidate pool (top 100 per query) is a viable substitute for full-corpus embedding, and what score fusion can extract from that restricted budget. We implement BM25@100 as the first stage on the MS MARCO document development collection (5193 topics), train a passage-level BERT bi-encoder with 24-dimensional compression, and evaluate four pool-level fusion strategies: semantic-only re-ranking, linear blending, rank-stratified blending, and reciprocal rank fusion. Linear blending at alpha = 0.85 (BM25-heavy) yields nDCG@10 of 0.3197 versus 0.3155 for BM25 alone (p = 0.012, paired bootstrap); semantic-only re-ranking falls to 0.0904 (-71% relative); the remaining variants trail BM25 by 8-11%. Under pool restriction, reducing the lexical weight consistently degrades effectiveness; the passage-trained encoder does not compensate for the lost BM25 signal at document granularity. Cross-encoder reranking and learned feature stacking are under ongoing investigation.

Keywords: information retrieval; MS MARCO; BM25; BERT; dense retrieval; score fusion; reciprocal rank fusion; document ranking; empirical evaluation

1. INTRODUCTION

Effective document retrieval at web scale generally requires a two-stage architecture: a fast first stage narrows a large corpus to a manageable candidate set, and a more expensive second stage re-scores those candidates using richer representations. Neural re-rankers based on pre-trained language models have become standard in this second stage (Lin, Ma, et al. 2021), but their application to long documents introduces practical constraints not present in passage-level tasks. Documents often exceed the token budget of a standard transformer, requiring segmentation strategies; encoding an entire corpus at dense precision is expensive both offline (embedding storage and computation) and online (vector search over millions of entries).

Restricting dense encoding to the first-stage candidate pool is one response to this cost. Rather than building a separate approximate nearest-neighbor index over all documents, the pool-restricted design embeds only those documents BM25 already returns, avoiding full-corpus infrastructure while concentrating compute on the candidates most likely to be relevant. The trade-off is that any relevant document outside the pool is permanently inaccessible to dense and fusion stages, so BM25 recall becomes a hard ceiling on the pipeline. This paper examines what score fusion can extract within that constraint. We implement BM25 as the first stage on the MS MARCO document development benchmark (5193 dev topics), train a passage-level bi-encoder and encode only the BM25@100 pool per query, and apply four fusion strategies. Macro-averaged ranking metrics with paired bootstrap confidence intervals are reported for each strategy. The experiments are designed to isolate where lexical scores dominate dense or fused alternatives under a fixed pool-restricted encoding budget.

The study does not include cross-encoder reranking or learned stacking; those represent ongoing work and will be addressed in subsequent publications.

2. RELATED WORK

2.1 Benchmarks and Evaluation

Craswell et al. (Craswell et al. 2021) describe the MS MARCO and TREC Deep Learning tracks as large-scale benchmarks connecting dense labels to pooled judgment infrastructure. The document development split used here provides 5193 queries with official qrels against a corpus of approximately 3.2 x 10^6 web documents. Arabzadeh et al. (Arabzadeh et al. 2022) analyse the effects of shallow pooling and sparse binary labels, noting that pooling depth and label format jointly determine which retrieval behaviors receive credit under offline metrics. MacAvaney et al. (MacAvaney, Macdonald, and Ounis 2022) provide the evaluation toolkit used throughout this study.

2.2 Lexical First Stages

BM25 (Robertson and Zaragoza 2009) remains a competitive first stage at depth 100 on many IR benchmarks despite the growth of neural alternatives. Ma et al. (Ma et al. 2022) demonstrate that reproducible sparse and dense hybrid baselines on MS MARCO V1 can match or approach higher-cost neural systems when first-stage recall is sufficiently high. Hofstätter et al. (Hofstätter et al. 2021) introduce an intra-document cascade that selects passages cheaply before expensive BERT scoring; the core idea of reducing encoder cost by restricting which documents receive dense representations is shared with the pool-restricted design studied here. The present study uses BM25 with Lucene parameters tuned for the document collection and does not apply query expansion or pseudo-relevance feedback.

2.3 Dense and Passage-Level Bi-Encoders

Wang et al. (Wang, Huang, and Sheng 2024) introduce a presentation-learning framework for long documents: a bi-encoder scores short passages and aggregates per-passage scores to a document-level relevance estimate via a max-over-segments operation. This architecture is adopted here with a 24-dimensional projection to reduce encoding cost. Nogueira and Cho (Nogueira and Cho 2019) establish BERT-based passage re-ranking as a strong two-stage baseline; the cross-encoder component is deferred to future work while this study focuses on the fusion tier. Huang et al. (Huang et al. 2023) study joint training of dense retrieval with product-quantized representations for latency-aware deployment, supporting the case for low-dimensional compression. Chen et al. (Chen et al. 2023) show that training-domain mismatch between passage triples and web documents can substantially reduce BERT re-ranking quality, a finding directly applicable to the passage-trained encoder used here.

2.4 Score Fusion

Reciprocal Rank Fusion (RRF) (Cormack, Clarke, and Buettcher 2009) is a parameter-light method for combining ranked lists that has shown competitive performance against learned fusion on several benchmarks. Linear score blending, in which normalized scores from two systems are added with a mixing parameter, is a common baseline for hybrid retrieval. Rank-stratified blending is a variant that applies different mixing weights to high-rank and low-rank candidates, motivated by the observation that lexical and dense models tend to disagree most in the tail of a candidate list. These evaluation caveats bear directly on interpreting the +1.3% nDCG@10 margin for linear blending: Lin et al. (Lin, Campos, et al. 2021) show that such margins on the MS MARCO document leaderboard frequently fail to replicate, and Gupta and MacAvaney (Gupta and MacAvaney 2022) identify survivorship bias in judgment pools that may further distort comparisons. Wu et al. (Wu et al. 2023) demonstrate that aggregate metric gains can coexist with vulnerability to minor query perturbations. Paired bootstrap intervals are reported throughout to give the uncertainty context these aggregate scores alone cannot supply.

3. BACKGROUND AND NOTATION

Let q denote a query, d a document, and Cq the BM25@100 candidate pool for q, with |Cq| = 100 in all experiments below. Relevance labels rel(q,d) come from official qrels. Rank position is 1-based; macro-averaged metrics average over all Q = 5193 dev topics.

3.1 BM25 Scoring

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages: 833-836

$$BM25(q,d) = \sum_{t \in q} w_t^{(q)} \frac{f_{t,d}(k_1+1)}{f_{t,d} + k_1(1 - b + b|d|/\text{ avgdl})'}$$

with inverse document frequency weight $w_t^{(q)}$, term frequency $f_{t,d}$, document length $|d|$, corpus average length avgdl, and parameters (k_1, b).

3.2 Bi-Encoder

Score .Let $h^q \in R^{L_q \times d}$ and $h^p \in R^{L_p \times d}$ be compressed BERT hidden states for a query and a single passage segment. The per-segment score via maxsimilarity alignment (Wang, Huang, and Sheng 2024) is

$$s_{\text{seg}}(q,p) = \cos\left(h^{-q}, \frac{1}{L_q}\sum_{j=1}^{L_q} h^p_{\text{argmaxcos}\left(h_j^q h_\ell^p\right)}\right),$$

where h^{-q} is the mean query-token embedding. The document score aggregates over non-overlapping segments: $s_{\text{sem}}(q,d) = \max_{\text{pesegs}(d)} s_{\text{seg}}(q,p)$.

3.3 Retrieval Metrics

$$BM25(q,d) = \sum_{t \in q} w_t^{(q)} \frac{f_{t,d}(k_1+1)}{f_{t,d} + k_1(1 - b + b|d|/\text{ avgdl})'}$$

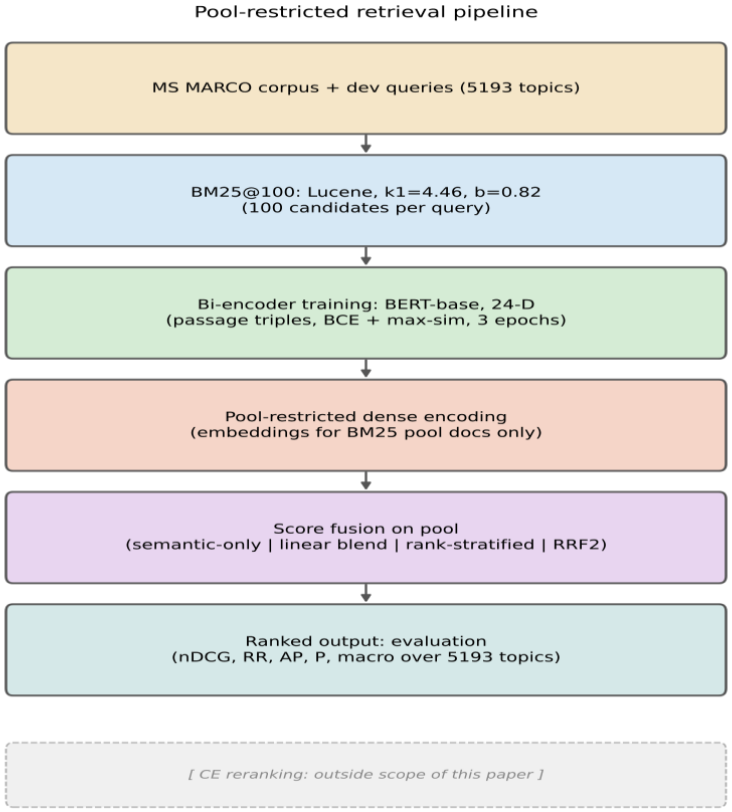
with inverse document frequency weight $w_t^{(q)}$, term frequency $f_{t,d}$, document length $|d|$, corpus average length avgdl, and parameters (k_1, b). Bi-Encoder Score .Let $h^q \in R^{L_q \times d}$ and $h^p \in R^{L_p \times d}$ be compressed BERT hidden states for a query and a single passage segment. The per-segment score via maxsimilarity alignment (Wang, Huang, and Sheng 2024) is

$$s_{\text{seg}}(q,p) = \cos\left(h^{-q}, \frac{1}{L_q}\sum_{j=1}^{L_q} h^p_{\text{argmaxcos}\left(h_j^q h_\ell^p\right)}\right),$$

where h^{-q} is the mean query-token embedding. The document score aggregates over non-overlapping segments: $s_{\text{sem}}(q,d) = \max_{\text{pesegs}(d)} s_{\text{seg}}(q,p)$.

4. METHODS

4.1 Experimental Pipeline



MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages: 833-836

Figure 1: Pool-restricted retrieval pipeline used in the experiment. Cross-encoder reranking is outside the scope of this paper.

4.2 Corpus and Queries

The MS MARCO document ranking collection (Craswell et al. 2019) contains approximately 3.2 x 10^6 web documents. We use the public development split (5193 topics) with official qrels. The bi-encoder is trained on the MS MARCO passage train-triples stream (approximately 4.0 x 10^7 triples in the standard "small" release), which is much larger in row count than the document corpus.

4.3 First Stage: BM25@100

We use PySerini's LuceneSearcher (Lin, Ma, et al. 2021) with parameters (k_1, b) = (4.46,0.82) and retrieve the top 100 documents per query with no query expansion. This produces the candidate pool Cq for all subsequent stages.

4.4 Bi-Encoder Training

Architecture: BERT-base (uncased) with a linear projection $R^{768} \rightarrow R^{24}$. Query tokens are prefixed with a [Q] marker; passage tokens with [D]. Query length cap $L_q = 50$; passage length cap $L_p = 150$ content wordpieces. Training minimizes a binary cross-entropy loss on passage triples:

$$L = BCEWithLogits\big(s_{seg}(q, p^*), 1\big) + BCEWithLogits\big(s_{seg}(q, p^-), 0\big),$$

using AdamW ($lr = 2 \times 10^{-5}$, weight decay 0.01, gradient clipping at 1.0) for three epochs with linear wamup. One checkpoint is used for all retrieval experiments.

4.5 Pool-Restricted Dense Encoding

Dense segment embeddings are computed only for documents whose IDs appear in at least one BM25@100 pool across the 5193 dev queries. Document bodies are truncated to 1000 characters before segmentation; segments are non-overlapping windows of up to Lp = 150 content tokens and stored in HDF5. Full-corpus encoding is not performed.

4.6 Score Normalization and Fusion

For each query's pool C_q, both BM25 and dense scores are independently min-max normalized to [0,1] :

$$s_d = \begin{cases} 0.5 s_{\max} = s_{\min}, \\ \dfrac{s_d - s_{\min}}{s_{\max} - s_{\min}} \text{ otherwise.} \end{cases}$$

Let $\hat{b}_d$ and $\hat{\delta}_d$ denote the normalized BM25 and semantic scores.

4.7 Fusion Strategies

Semantic-only.

Re-rank C_q by δ_d alone. This measures the performance of the dense model when it has no lexical signal.

Linear α-blend.

$$s_d^{\text{lin}} = \alpha \hat{b}_d + (1 - \alpha)\delta_d, \alpha = 0.85$$

The weight $\alpha = 0.85$ is fixed and gives a predominantly lexical score with a small dense component.

Rank-stratified blend.

Let $r^{BM25}(d)$ be the BM25 rank within C_{q^*}. With cutoff $r^* = 50$ and weights $\alpha_{\text{head}} = 0.72, \alpha_{\text{tail}} = 0.35$:

$$\alpha(d) = \{\alpha_{\text{head}}\, r^{\text{BM 25}}(d) \leq r^*, \alpha_{\text{tail}} \text{ otherwise, } s_d^{\text{strat}} = \alpha(d) b_d + (1 - \alpha(d))\delta_d.$$

Reciprocal Rank Fusion (RRF2).

Fuse BM25 and semantic-only orderings using Eq. [eq:nf] with $m = 2$, weights $(w_{\text{BMZS}}, w_{\text{sem}}) = (1.5,0.5)$, and $K = 80$.

5. RESULTS

Table 1: Macro-averaged retrieval metrics on MS MARCO document dev (5193 topics). All methods operate within the BM25@100 pool. Bold indicates best value per column.

Method	nDCG@10	nDCG@100	RR@10	AP@100	P@10
BM25	0.3155	0.3773	0.2565	0.2690	0.0506
Semantic-only	0.0904	0.2081	0.0653	0.0834	0.0173
Linear blend (alpha = 0.85)	**0.3197**	**0.3793**	**0.2587**	**0.2711**	**0.0515**
Rank-stratified	0.2889	0.3606	0.2411	0.2541	0.0442
RRF2 (BM25+semantic)	0.2812	0.3428	0.2121	0.2245	0.0507

Linear blending improves nDCG@10 by +1.3% over BM25 (absolute: 0.3155 -> 0.3197). Paired bootstrap over 5193 topics assigns p = 0.012 for this difference on nDCG@10; the differences on nDCG@100, RR@10, and AP@100 are not significant at alpha = 0.05. Semantic-only re-ranking trails BM25 by a large margin on all headline metrics (-71% on nDCG@10 relative), confirming that the passage-trained encoder does not transfer well to document-level ranking without lexical support. Rank-stratified blending and RRF2 both trail BM25 on nDCG@10 and nDCG@100, with rank-stratified slightly outperforming RRF2 on most columns.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages: 833-836

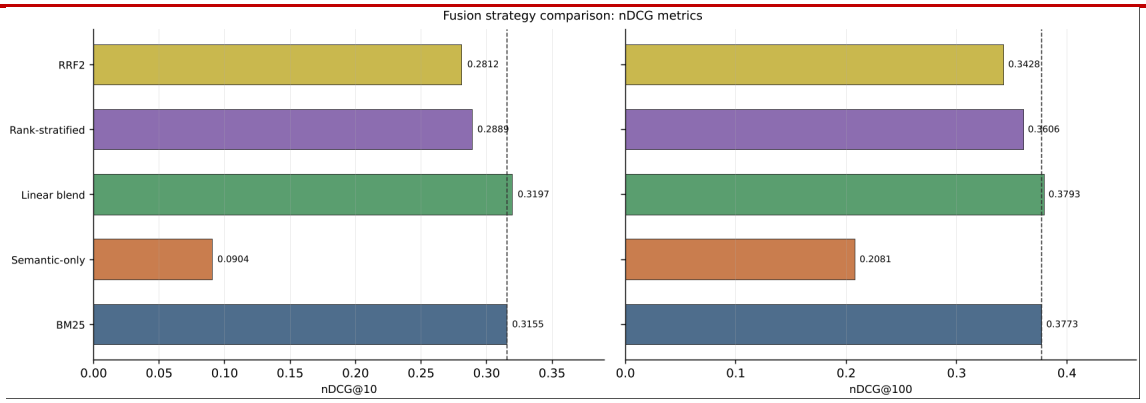


Figure 2: nDCG@10 and nDCG@100 for all five systems. BM25 is shown as a dashed reference line on each panel.

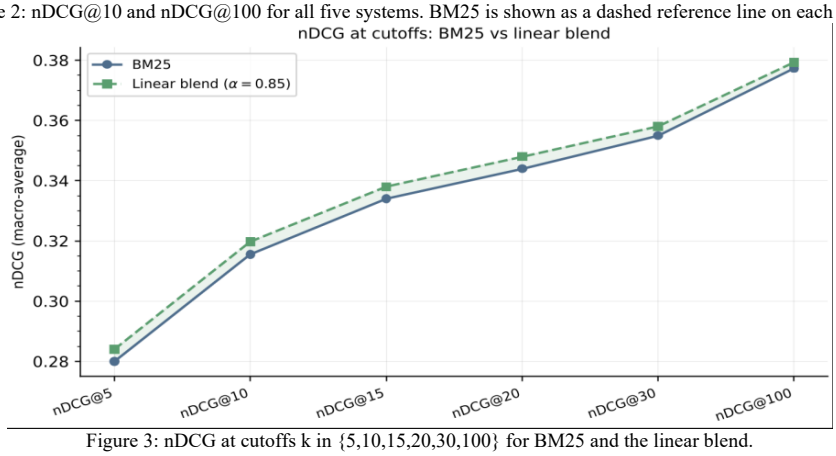


Figure 3: nDCG at cutoffs k in {5,10,15,20,30,100} for BM25 and the linear blend.

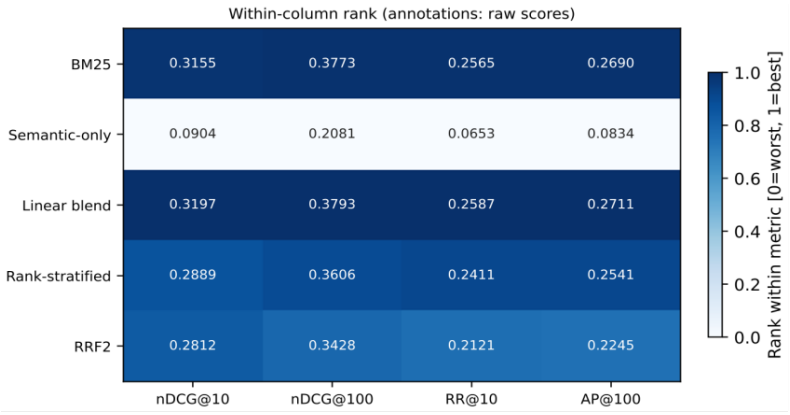


Figure 4: Headline metrics heatmap for all five systems. Color encodes within-column rank and annotations show raw scores.

6. DISCUSSION

Pool-restricted encoding produces a consistent pattern across all four fusion variants: lexical weight determines rank quality, and reducing it hurts. The bi-encoder was trained on short passage triples; document scores are then derived by taking the maximum cosine over non-overlapping segments. Max-over-segments aggregation loses global document context and produces noisy estimates for long documents with mixed content.

All dense scores are computed inside the BM25 pool, so their distribution is conditioned on BM25 selection. Reversing the BM25 ordering among 100 pre-selected candidates using a passage-level signal is harder than re-ranking from a full-corpus dense index where the embedding space was shaped on the full distribution.

Fusion strategies that lower the lexical weight cause large metric drops. The dense model improves precision within the top BM25 ranks but cannot compensate for lost lexical ordering across the full pool. All metrics use the official MS MARCO document development qrels. Relevant documents outside the BM25@100 pool are invisible to every stage and cannot contribute to recall-oriented metrics.

7. CONCLUSION AND FUTURE WORK

Four score fusion strategies were evaluated under a pool-restricted dense encoding constraint on the MS MARCO document development benchmark. Linear score blending (alpha = 0.85) produces a small but statistically significant nDCG@10 improvement over BM25 ($p = 0.012$); the three remaining variants all fall below the lexical baseline. Within a BM25@100 pool, a passage-trained BERT encoder at 24-dimensional compression cannot recover ranking quality when lexical weight is reduced.

Future work will examine cross-encoder reranking on short prefixes of the candidate list and learned stacking of lexical and dense features via out-of-fold logistic fusion. Extending the pool size and comparing pool-restricted encoding against full-corpus dense indexing are additional directions under investigation.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages: 833-836

8. DECLARATIONS

Funding: No specific funding was received for this work.
Conflict of interest: The authors declare no competing interests.
Data availability: The MS MARCO document ranking data, development queries, and qrels are publicly available from the dataset distributors.
Author contributions: S. M. Quadri: experiments, analysis, writing.

9. REFERENCES

Arabzadeh, Negar, Alexandra Vtyurina, Xinyi Yan, and Charles L. A. Clarke. 2022. "Shallow Pooling for Sparse Labels." Information Retrieval Journal 25(4): 365-385. https://doi.org/10.1007/s10791-022-09411-0.

Chen, Xuanang, Ben He, Kai Hui, Le Sun, and Yingfei Sun. 2023. "Dealing with Textual Noise for Robust and Effective BERT Re-Ranking." Information Processing & Management 60(1): 103135. https://doi.org/10.1016/j.ipm.2022.103135.

Cormack, Gordon V., Charles L. A. Clarke, and Stefan Buettcher. 2009. "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods." In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 758-759. Boston, MA: ACM. https://doi.org/10.1145/1543834.1543934.

Craswell, Nick, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. "MS MARCO: Benchmarking Ranking Models in the Large-Data Regime." In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1566-1576. https://doi.org/10.1145/3404835.3462804.

Craswell, Nick, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2019. "Overview of the TREC 2019 Deep Learning Track." Text REtrieval Conference (TREC 2019), National Institute of Standards and Technology. https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.DL.pdf.

Gupta, Prashansa, and Sean MacAvaney. 2022. "On Survivorship Bias in MS MARCO." In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2214-2219. https://doi.org/10.1145/3477495.3531832.

Hofstätter, Sebastian, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021. "Intra-Document Cascading: Learning to Select Passages for Neural Document Ranking." In Proceedings of the 44th International ACM SIGIR Conference, 1349-1358. https://doi.org/10.1145/3404835.3462889.

Huang, Rong, Danfeng Zhang, Weixue Lu, Han Li, Meng Wang, Daiting Shi, Jun Fan, Zhicong Cheng, Simiu Gu, and Dawei Yin. 2023. "Learning Discrete Document Representations in Web Search." In Proceedings of the 29th ACM SIGKDD Conference, 4185-4194. https://doi.org/10.1145/3580305.3599854.

Lin, Jimmy, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. 2021. "Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard." In Proceedings of SIGIR 2021, 2283-2287. https://doi.org/10.1145/3404835.3463034.

Lin, Jimmy, Xueguang Ma, Sheng-Chieh Lin, Ji-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations." In Proceedings of SIGIR 2021, 2356-2362. https://doi.org/10.1145/3404835.3462985.

Ma, Xueguang, Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2022. "Document Expansion Baselines and Learned Sparse Lexical Representations for MS MARCO V1 and V2." In Proceedings of SIGIR 2022, 3187-3197. https://doi.org/10.1145/3477495.3531749.

MacAvaney, Sean, Craig Macdonald, and Iadh Ounis. 2022. "Streamlining Evaluation with Ir-Measures." In Advances in Information Retrieval: ECIR 2022, 309-313. https://doi.org/10.1007/978-3-030-99736-6_24.

Nogueira, Rodrigo, and Kyunghyun Cho. 2019. "Passage Re-Ranking with BERT." arXiv preprint arXiv:1901.04085. https://arxiv.org/abs/1901.04085.

Robertson, Stephen, and Hugo Zaragoza. 2009. "The Probabilistic Relevance Framework: BM25 and Beyond." Foundations and Trends in Information Retrieval 3(4): 333-389. https://doi.org/10.1561/1500000019.

Wang, Junmei, Jimmy X. Huang, and Jinhua Sheng. 2024. "An Efficient Long-Text Semantic Retrieval Approach via Utilizing Presentation Learning on Short-Text." Complex & Intelligent Systems 10: 963-979. https://doi.org/10.1007/s40747-023-01192-3.

Wu, Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2023. "Are Neural Ranking Models Robust?" ACM Transactions on Information Systems 41(2): 1-36. https://doi.org/10.1145/3534928.