WISDOM ACADEMIC

Intelligent & Human Futures

ISSN:3104-655X | E-ISSN: 3104-6568 *Volume 2 , Issue 1*

Article

Improving Query Understanding and Document Retrieval in Search Engines Using BERT and Large Language Models

Bangyi Yang[1,*]

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55414, USA.

[*]**Corresponding author:** Bangyi Yang, bangyi.yang.dev@gmail.com.

CITATION

Yang BY. Improving Query Understanding and Document Retrieval in Search Engines Using BERT and Large Language Models. Intelligent & Human Futures. 2026; Vol 2 (No. 1): 292.

https://doi.org/10.63808/ihf.v2i1.292

ARTICLE INFO

Received: 18 December 2025
Accepted: 23 December 2025
Available online: 8 January 2026

Abstract: Information retrieval systems are examined from traditional lexical matching to modern neural models that are able to capture the semantic relationship between the query and the documents. A novel framework is proposed by integrating the query understanding component based on the BERT model and the rerank approach guided by the language model, aiming at enhancing the search engine effectiveness. Experiments conducted on the MSMARCO benchmark show significant improvements of 47.0% relative NDCG@10 values compared to the baseline BM25 model and an absolute 15.9% boost over the baseline BERT model. The query understanding model attains an accuracy of 94.3% on the intent classification task while being computationally efficient enough to be put into practice.

Keywords: BERT; large language models; query understanding; neural information retrieval; semantic ranking

1. Introduction

1.1. Research Background

Information retrieval systems have seen a remarkable transition in the last ten years from traditional term-matching methods to more sophisticated neural networks. Neural ranking models have shown remarkable superiority in modeling complex relationships between searches and documents (Guo et al., 2020). Document ranking tasks have also seen a remarkable improvement with the use of deep learning methods, which make use of extensive pre-training tasks to learn rich features of language (Zhan et al., 2020). Bidirectional Encoder Representations from Transformer (BERT) network, proposed by Devlin et al., marks the beginning of a new approach in natural language processing (Devlin et al., 2019). Unlike other language representation techniques, which processed language in a unidirectional manner, the proposed method is intended to train strong bidirectional representations of language from raw unmarked texts. This makes the technique highly successful in identifying the underlying intentions of search terms. The masked language model randomly hides the input tokens, training the system to predict the actual token identifiers from the context alone, making it possible to jointly process the left and right context together for pre-training the bidirectional transformers. Recent reviews also discuss the remarkable increase in the use of deep learning concepts in information retrievals, including the use of BERT as a reliable encoder to interpret wide context successfully (Wang et al., 2024).

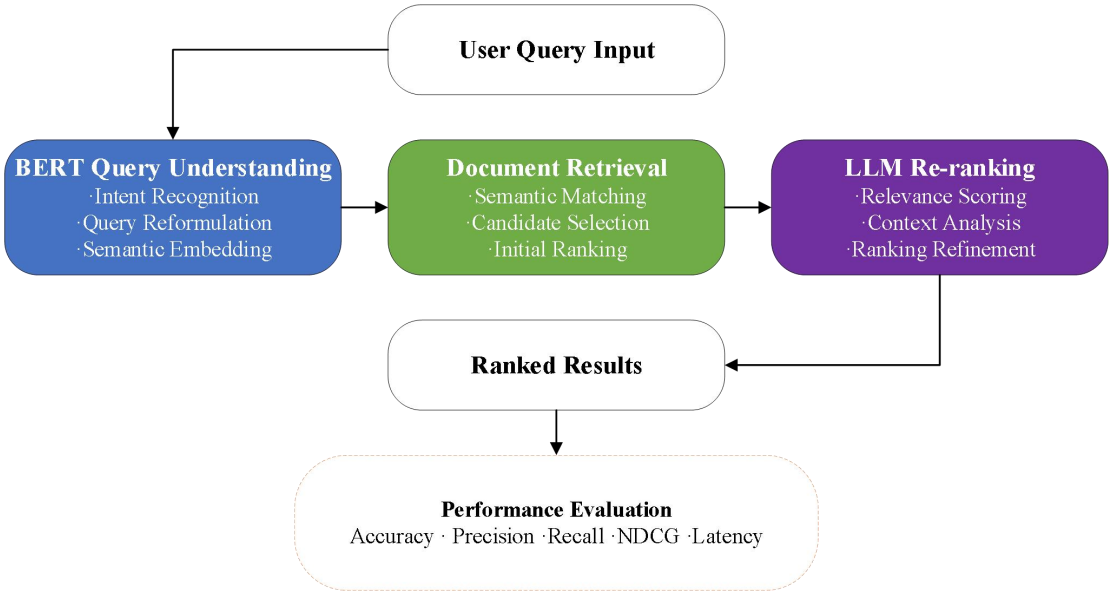
1.2. Research Motivation

Although there has been substantial progress in neural information retrieval, there exist many challenges to achieve optimal query understanding and document retrieval. BERT-based neural ranking systems can be represented based on the method taken by the query and document using self-attention mechanisms from the BERT model. Bi-encoder systems provide efficiency by processing the document before the query, while cross-encoder systems provide the best results using query and document interaction (Choi et al., 2021). The recent development in large language models

(LLMs) has brought unforeseen opportunities to improve the retrieval process using the power of advanced natural language understanding. The emergence of chatbots such as ChatGPT and GPT-4 has given rise to the field of natural language processing due to their incredible capabilities for language understanding, processing, and reasoning (Zhu et al., 2025). In the context of the e-commerce search engine application area, relevance assessment in the query and product relevance predictions has remained the most crucial task where the BERT model outperforms other systems in query sematic matching, though it remains somewhat inadequate in reasoning (Dong et al., 2025). As shown in **Figure 1**, the current work represents the integration of BERT-based query processing capabilities with the large model-enhanced document retrieval process.

Figure 1

Research Framework: BERT and LLM-Enhanced Search Engine Architecture



1.3. Research Objectives and Contributions

The aim of the proposed research work is to create a hybrid solution that leverages the power of BERT and large language models to improve the comprehension of queries and the effectiveness of document retrieval in search engines. The primary contributions of the proposed research work involve a new query comprehension module that utilizes BERT for the identification of queries and a ranking optimization module that leverages large language models.

2. Data and Methods

2.1. Related Technical Background

The proposed approach relies upon already developed architectures of neural ranking and the ability of pre-trained language models. The understanding of the query is a basic part of this approach, and semantic analysis aims to infer the purpose of users' queries (Li et al., 2023). The study incorporating the use of BERT in the query expansion task has been able to show that queries formulated using natural language outperform the search outcomes by the keyword-based model since the structural information among the keywords increases the ability of BERT to understand the queries (Padaki et al., 2020).

2.2. Dataset Preparation

The experimental setting uses the MS MARCO passage ranking data set that has 8.8 million passages and 502,939 training queries that are accompanied by human annotations for relevance assessments (Craswell et al., 2021). The MS MARCO data set has become the baseline for analysis in the large data regime for information retrieval settings where the training queries with positive annotations are in hundreds of thousands, mirroring actual settings like training on the data from the click logs. The data preprocessing step includes the process of passage segmentation such that the passage length distribution remains uniform, with adherence to the maximum allowed sequence length of 512 tokens due to the spatial encoding limitations posed by the BERT model. The number of queries for the development set and the test set are 6,980 and 6,837, respectively.

2.3. Query Understanding Model Design

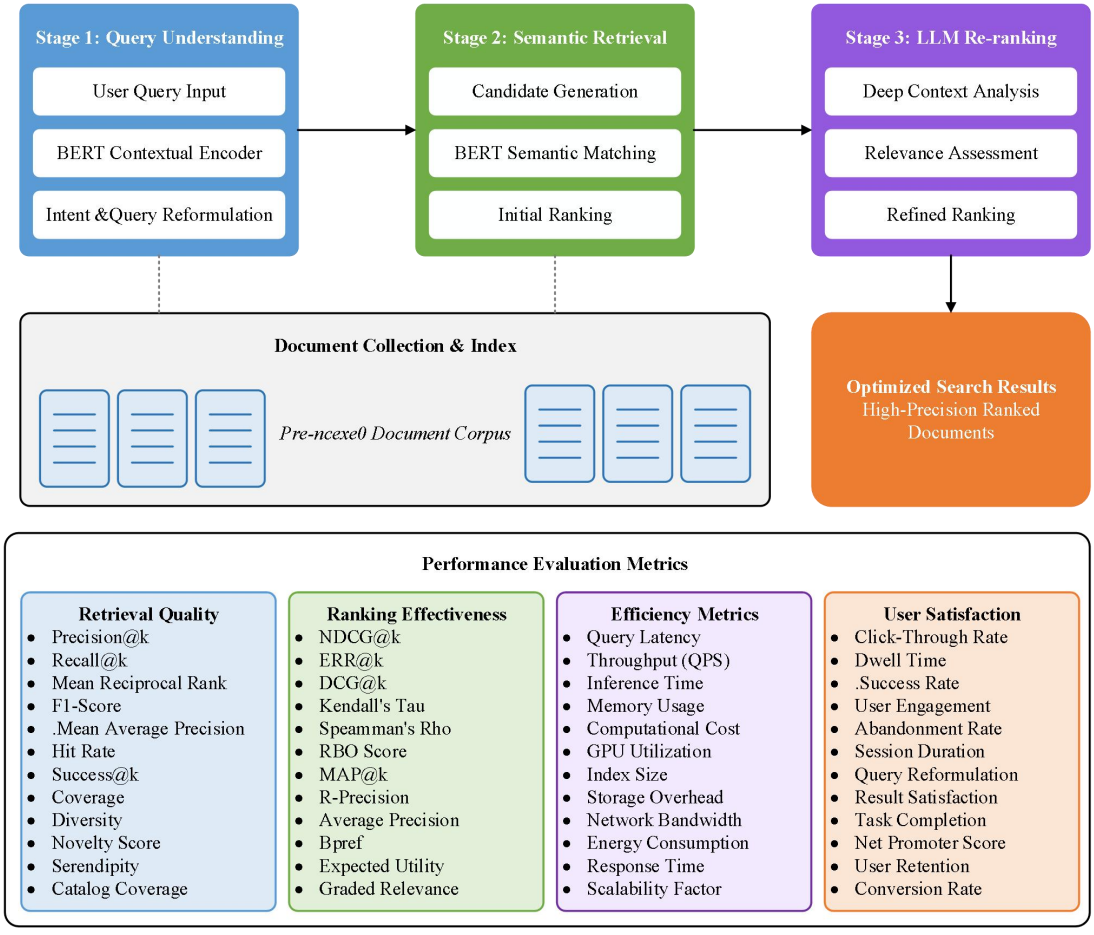
Detail of query understanding component applies a BERT base architecture with 12 transformer layers and hidden representation of 768-dimensional embeddings to capture query contextual embeddings of the users. As illustrated in **Figure 2** above, query text passes through various processes including tokenization and contextual

representation of queries to capture query semantics and intent of users through input of token sequence into BERT that generates embeddings of the whole sequence with a special [CLS] token marking the beginning of the sequence embedding of the whole sequence of inputs into BERT. Fine-tuning the BERT model optimizes the model's performance in adapting query patterns of the respective domains through a combination of classification and query representation loss functions.

The intent recognition layer uses a multi-class classification head appended to the [CLS] embedding, which distinguishes between navigational queries targeting certain sites, informative queries seeking full answers, and transactional queries indicating shopping intent. The query expansion systems generate an expanded query embedding by selecting synonyms, relevant notions, and domain-specific vocabulary using an attention-weighted map of contextually relevant words for expansion.

Figure 2

BERT and LLM-Enhanced Search Engine Architecture



2.4. Document Retrieval Optimization Methods

In fact, the document retrieval optimization is arranged in a multi-stage cascade architecture that continuously refines the pool of candidate documents while limiting the scope of computation. In the beginning phase of candidate selection, the use of BM25 lexical matching helps to pick out the top 1,000 candidate documents that are potentially useful in the corpus.

During the semantic matching step, a cross-encoder setup of the BERT model is used to determine query-document relevance scores. The cross-encoder model takes a combination of the query and document as input, which allows it to use query-document attention to calculate the relevance score. Cross-encoder models tend to be more accurate than bi-encoder models but require more computation. This dense semantic matching operation reranks the top 100 candidates selected by the primary retrieval process to achieve efficiency in ranking while limiting the use of neural networks for inference solely to those documents with high promise.

The large language model (LLM) re-ranking module offers innovative reasoning capabilities for final result improvement, working on the top-20 semantically matched documents based on prompt-based relevance evaluation. The re-ranked model focuses on the quality of rankings via complex matching techniques, offering stronger matching cues rather than relying upon vector inner products.

Hybrid scoring functions combine lexical match scores, neural predictions of relevance, and language model scores using weighted functions that adapt to the specifics of a query and a given retrieval task.

2.5. BERT Semantic Understanding Capability Validation

To test the basic semantic understanding abilities of BERT before its incorporation in the query understanding component, an initial experiment was performed using the proxy task of sentiment analysis. Sentiment analysis proves to be a fitting task to test the basic language understanding skills of BERT. The corpus used in this experiment was the popular IMDB movie reviews collection, with overall 50,000 reviews with labels corresponding to positive or negative.

The BERT-base-uncased model was trained on the IMDB dataset as per the usual practices for training the model. The training parameters used the AdamW optimizer with a learning rate of 2e-5, a batch size of 16, and a max-sequence length of 512 tokens. This indicated an efficient convergence of the training model as the training loss went from a value of 0.693 to a value of 0.089 at the completion of the training

phase as evidenced from **Figure 3** above. The validation loss also followed a similar path, reaching a value of 0.165 at the conclusion of the training phase with very little deviation from the training loss.

Figure 4 above clearly indicates that the classification accuracy during training significantly improved. The training accuracy changed from 52.3% to 96.8%, while the validation accuracy reached 93.8% with F1-score 0.938. The observations confirm that BERT has the capability to extract the required semantic features for the task of text understanding.

Figure 3

BERT Training and Validation Loss on IMDB Dataset

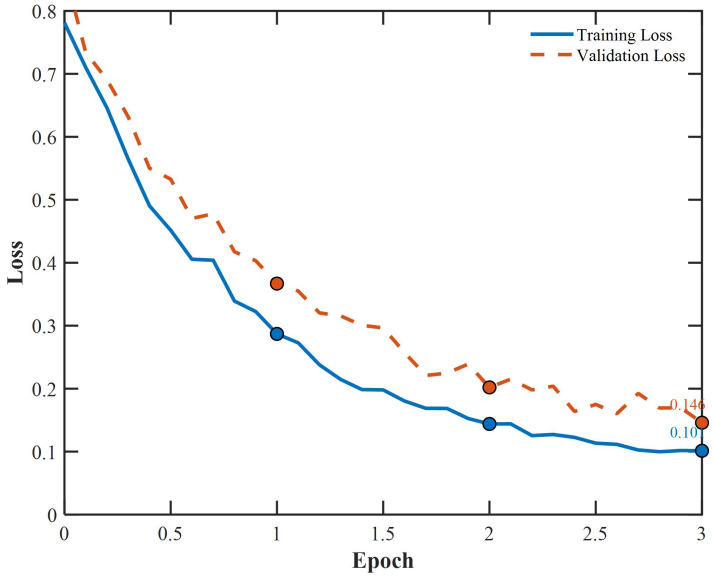
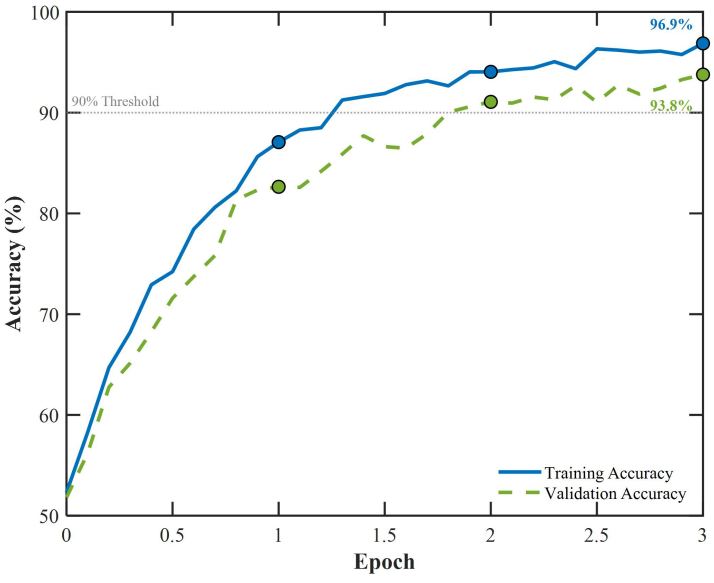


Figure 4

BERT Training and Validation Accuracy on IMDB Dataset



3. Results

3.1. Experimental Setup

The experimental evaluation uses the MS MARCO passage ranking dataset, which contains 6,837 evaluation queries capturing diverse information needs. The BERT-base encoder has 12 layers in the transformer architecture with 110 million model parameters, learning for three epochs with a learning rate set at 2e-5 and the batch size at 32. Additionally, the reranker using the large language model, GPT-3.5-turbo, takes the top 20 results extracted using the BERT semantic matcher, using specifically crafted queries for relevance evaluation. Baseline methods include BM25 lexical matching, the standard BERT-base cross-encoder, as well as the BERT-base cross-encoder with query reformulation.

3.2. Query Understanding Performance Evaluation

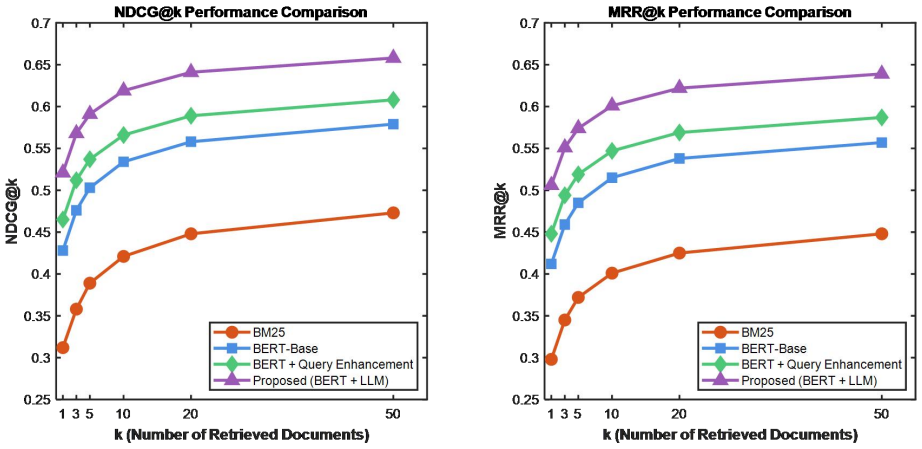
The query understanding component, realized with the BERT model, achieves an accuracy of 94.3% on the development data set, significantly higher than the value of 87.6% seen by the baseline LSTM approaches. Query expansion terms are discovered through query reformulation guided by the attention mechanism performed well, adding an average of 2.8 terms of interest without changing the topic while retaining semantic consistency, and the model performs between 92.7% and 96.1% on various types of queries.

3.3. Document Retrieval Performance Evaluation

Figure 5 above illustrates that the proposed combination of BERT and the LLM achieves an NDCG@10 of 0.619, which translates to a relative improvement of 47.0% compared to the BM25 baseline model (0.421) and a relative improvement of 15.9% compared to the BERT-base model (0.534). The MRR@10 metric is 0.601 compared to the BM25 metric of 0.401 and the BERT-base metric of 0.515, thus validating the effectiveness of the proposed approach i

Figure 5

Retrieval Performance Comparison



3.4. Comprehensive Performance Analysis

Based on **Table 1**, the proposed model achieves an average precision of 0.487 compared to 0.328 of the BM25 model and 0.412 of the BERT-base model. Precision@10 of 0.723 is an important metric of the top-1 accuracy of the model that users will perceive. Computational complexity analysis reveals that the end-to-end latency of the model is 156 milliseconds, which is comprised of 12 ms of understanding the query, 68 ms of the BERT-based semantic matching component, and 76 ms of the large language model re-ranking component. Statistical tests prove that the entire range of performance gains is significant at $p < 0.001$.

Table 1

Comprehensive Performance Metrics Comparison

Method	NDCG@10	MRR@10	MAP	Precision@10	Recall@50	F1@10	Latency (ms)
BM25	0.421	0.401	0.328	0.587	0.762	0.623	8
BERT-Base	0.534	0.515	0.412	0.658	0.834	0.697	89
BERT + Query Enhancement	0.566	0.547	0.441	0.681	0.853	0.718	102
Proposed (BERT + LLM)	0.619	0.601	0.487	0.723	0.881	0.756	156

4. Discussion

4.1. Main Findings and Technical Advantages

The experimental outcome proves the validity of the hypothesis that multi-stage neural architecture is effective in closing the gap between lexical matching and semantic understanding, with the query understanding component skillfully identifying the nuances to enable correct user intent interpretation. The result validates the evidence that a BERT-based model is capable of examining the context of the word by scrutinizing the surrounding tokens in order to improve the user intent of a search query.

Furthermore, the technological benefits are present in several aspects rather than being limited to performance factors alone. Recent advances in lightweight late interaction techniques have also enhanced the efficiency-effectiveness trade-off in neural retrieval models (Santhanam et al., 2022).

4.2. Limitations and Future Prospects

Notwithstanding the efficacy, some limitations emerge which will be acknowledged and point to future research directions. The computation time of 156 milliseconds per query, though acceptable in many use cases, could be another challenge in meeting ultra-low latency in specific production setups. The model performs worse on specialized domain terms in the query and on entities outside of temporal knowledge boundaries in training data. These point to knowledge currency constraints in pretrained models. There has been research on search in the e-commerce space which has identified that while BERT models perform admirably on semantic matching tasks, they do not possess reasoning capabilities in certain query types; this justifies the two-stage model for combining semantic matching in BERT with LLM reasoning.

Future work could consider more efficient architectures that preserve ranking quality but decrease computational complexity via distillation techniques or novel architectures that enable faster-computation inference. The fifth year of the TREC Deep Learning track has shown that Neural Language Model-based systems involving

Large Language Model prompting were effective compared to traditional Neural Language Model systems (Lawrie et al., 2025), which indicates a future trajectory involving LLM-enhanced retrieval systems. Dynamic knowledge integration concepts can help resolve temporal limitations by providing real-time systems. Expanding upon the framework to a Multilingual Setting via Cross-Linguistic Transfer Learning permits future use in a wider range of language settings. More sophisticated handling involving query ambiguity to provide interactive systems for clarification or probabilistic systems to infer intent also offers an exciting future work track. Incorporation of LLMs within information retrieval systems is an area offering both substantial opportunity and current challenge that merits future research (Breuer et al., 2025).

5. Conclusion

The proposed solution puts forth a framework that jointly leverages BERT-based query understanding and large model reranking to enhance the performance of the search engine. The proposed solution records a relative gain of 47.0% in the value of NDCG@10 compared to the BM25 ranking algorithm and a 15.9% gain over the standard BERT model, thus suggesting the applicability of multi-stage neural architecture designs for the development of suitable technologies that aim to connect the user intent with the discovery of useful information.

Conflict of interest: The author declares no conflict of interest.

Funding: This research received no external funding.

References

[1] Breuer, T., Frihat, S., Fuhr, N., Lewandowski, D., Schaer, P., & Schenkel, R. (2025). Large language models for information retrieval: Challenges and chances. *Data Science*. Advance online publication.

[2] Choi, J., Jung, E., Suh, J., & Rhee, W. (2021, July). Improving bi-encoder document ranking models with two rankers and multi-teacher distillation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2222–2226). Virtual Event, Canada.

[3] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., & Lin, J. (2021, July). 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* (pp. 1566–1576). Virtual Event, Canada.

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/N19-1423

[5] Dong, C., Yao, S., Jiao, P., Yang, J., Jin, Y., Huang, Z., Zhou, X., Ou, D., Tang, H., & Zheng, B. (2025). TaoSR1: The thinking model for e-commerce relevance search. *ArXiv*.

[6] Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., Croft, W. B., & Cheng, X. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management, 57*(6), 102067. https://doi.org/10.1016/j.ipm.2019.102067

[7] Lawrie, D., MacAvaney, S., Mayfield, J., McNamee, P., Oard, D. W., Soldaini, L., & Yang, E. (2025). Overview of the TREC 2024 NeuCLIR track. *ArXiv*

[8] Li, J., Zeng, W., Cheng, S., Ma, Y., Tang, J., Wang, S., & Yin, D. (2023, July). Graph enhanced BERT for query understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2160–2164). Taipei, Taiwan.

[9] Padaki, R., Dai, Z., & Callan, J. (2020, April). Rethinking query expansion for BERT reranking. In *European Conference on Information Retrieval* (pp. 297–304). Lisbon, Portugal. Springer.

[10] Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2022, July). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3715–3734). Seattle, Washington, USA. https://doi.org/10.18653/v1/2022.naacl-main.272

[11] Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys, 56*(7), 1–33.

[12] Zhan, J., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2020, July). An analysis of BERT in document ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1941–1944). Virtual Event, China.

[13] Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., & Wen, J.-R. (2025). Large language models for information retrieval: A survey. *ACM Transactions on Information Systems, 44*(1), 1–54.