Journal of Intelligent Information and Communication

Research Paper

Dual-Encoder BERT Approach for Cross-Domain Scientific Literature Retrieval

Marko Jovanović[1, *], Jelena Nikolić[2], Aleksandar Popović[2] and Dušan Đorđević[3]

[1] Faculty of Applied Information Technology, Belgrade Higher Professional School, 11070 Belgrade, Serbia
[2] Faculty of Information Technology, Alfa BK University, 11000 Belgrade, Serbia
[3] Faculty of Computer Science, University of Novi Sad, 21000 Novi Sad, Serbia
*Corresponding author: m.jovanovic@bpu.edu.rs

Abstract. Deep learning-based retrieval models perform well in large-scale scientific papers; however, due to domain and semantic differences, retrieving relevant literature from different knowledge domains remains a technical challenge. This paper introduces a BERT-based dual-encoder architecture for fast and accurate cross-domain scientific literature retrieval. The BERT encoder uses parameter sharing to encode queries and candidate documents, aligning their semantic representations in a shared embedding space. Throughout the entire experiment, the three heterogeneous scientific corpora covered over 4.2 million documents across twelve research fields. Using the BERT tokenizer, each text is normalized and processed, and then embedded into a vectorized index in FAISS for fast nearest neighbor search. Training uses batch InfoNCE loss and hard negative sampling, along with dynamic batch adjustment and early stopping mechanisms. According to the above empirical results, the proposed method achieves an average accuracy of 0.624, outperforming strong neural and traditional baselines such as ColBERT and BM25. In high-resource domains, the recall rate of the top ten exceeds 0.83, while it remains stable in low-resource domains, indicating its broad applicability. Ablation studies also indicate that batch-based negative sample mining and attention regularization require good performance; engineering analysis has already achieved efficient indexing and query latencies below 100 milliseconds. Based on the above findings, an interdisciplinary academic search engine can be constructed using a dual-encoder BERT model optimized with contrastive learning and scalable vector indexing.

Keywords: *Neural Information Retrieval, Cross-Domain Search, BERT, Dual-Encoder Architecture, Contrastive Learning, Scientific Document Indexing, Scalability, Semantic Embedding*

Received on 19 September 2025, Accepted on 06 January 2026, Published on 20 January 2026

Introduction

The expansion of science and the spread of knowledge have gradually changed in recent years [1]. Now, there is a large number of papers from many places and fields of science that researchers need to read [2]. With the development of new research trends, many fields are now crossing over, so we need better tools that can find related information in many areas simultaneously [3]. Traditional information retrieval (IR) systems have mainly been using keyword matching and shallow syntactic cues, and thus have failed to address the semantic and terminological discrepancies among different disciplines [4]. As a result, the problem of retrieving scientific literature from all fields is still difficult, and therefore simple string-based methods have low recall and precision in handling heterogeneous vocabularies and conceptual systems [5]. In addition, when the source of the query is different from the domain of the work it seeks, the semantic gap is more pronounced, and traditional IR pipelines are not suitable [6]. Many studies have shown that such pipelines are unable to model the complex contextual relations in scientific texts due to variations in the structure and style of different fields [7]. To address the above deficiencies and promote the efficient transfer of interdisciplinary knowledge for scientific innovation [8].

Deep learning and natural language processing have recently made significant progress, allowing text to be encoded into dense semantic representations through neural retrieval architectures [9]. These advancements

are due to the aforementioned issues. BERT is a context-aware language model that improves retrieval accuracy by learning context-sensitive representations [10]. However, most BERT-based retrieval frameworks are currently built using single-domain corpora, resulting in lower performance and limited generalization capabilities in cross-domain environments [11]. Due to their computational efficiency and suitability for large-scale matching, dual-encoder models have recently garnered attention, with these models including independent encoding of queries and documents [12]. Although the dual-encoder architecture can achieve real-time retrieval and scale to millions of corpora, it still faces the challenge of needing to be widely applied across a broad range of scientific fields [13]. Moreover, due to the significant differences in scientific concepts, terminology, and document structures, the domain transfer problem remains unresolved [14]. In the field of cross-domain scientific retrieval, effective training methods, domain adaptation mechanisms, and robust evaluation procedures remain the focus of research [15]. Benchmark studies have recently pointed out that retrieval models must effectively handle differences in language systems and concepts, as well as differences between disciplines, while maintaining high accuracy and low latency [16]. System scalability and cross-domain robustness are currently key research topics for the next generation of scientific search engines [17]. The integrated strategy addresses the aforementioned issues through advancements in the dual-encoder transformer model [18].

This paper proposes a dual-encoder BERT model for cross-domain scientific literature retrieval. By using powerful semantic encoding and contrastive learning, our framework can map query-document pairs into a shared vector space. It has extremely high generalizability and accuracy across various scientific fields. Based on the aforementioned experimental data, multiple heterogeneous benchmark sets were used to test the performance and generalizability of the new method. The main contributions are as follows: (1) We introduce a scalable dual-encoder BERT framework for robust cross-domain scientific retrieval; (2) We provide comprehensive quantitative and qualitative analyzes, including error analysis and ablation studies; (3) We discuss the system's computational efficiency and practical deployment scenarios. The following organizes the other sections of this paper: Section 2 introduces related research, Section 3 presents the dual-encoder BERT framework, Section 4 showcases the experimental results and analysis, and Section 5 concludes the paper and proposes future research directions.

Related Work

Cross-Domain Literature Retrieval

With the progress of interdisciplinary research in recent years, more and more scholars have begun to focus on cross-domain retrieval of scientific literature. Due to the characteristics of term matching and manual design, early literature retrieval systems were unable to understand the semantics of different scientific fields. The accuracy and recall rates of the aforementioned methods are generally low, especially when it comes to differences in terms and concepts across different fields [20]. Due to the complexity and error-proneness of mapping scientific terms between these fields, attempts to bridge the differences through external ontologies or lexicons have only made negligible progress [21].

In document retrieval within heterogeneous collections, unsupervised and semi-supervised topic modeling frameworks, such as Latent Dirichlet Allocation and its variants, have been recently used. However, their bag-of-words model characteristics are not suitable for complex scientific texts [22]. Cross-domain adaptation methods have been used for domain adversarial training and importance weighting to address domain-specific biases and improve retrieval accuracy; however, they remain relatively sensitive to significant changes in document structure and discourse [23]. Moreover, most classical methods are based on homogeneous benchmarks and do not consider the specific issues of multi-domain scientific corpora [24].

With the development of large-scale open-access digital libraries, these libraries have added a significant number of metadata to build retrieval models that consider citation graphs, collaboration networks, and other document attributes. Graph and network-centric methods are suitable for finding paths of knowledge transfer. However, they are costly and may not be suitable for real-time searches [25]. Due to the shortcomings of these metadata-driven traditional methods, new representation and computation methods are needed to meet the current cross-domain scientific search requirements [26].

Neural Information Retrieval Models

Using neural ranking models for information retrieval can more accurately match semantics, rather than just using word alignment [27]. Convolutional and recurrent neural networks were used to create documents and queries, as well as the first neural information retrieval system that learns relevance signals directly from raw text [28]. These models have achieved success in news and web search tasks, but due to the use of different terminologies and structures in academic writing, they perform poorly in scientific literature [29].

Dense vector retrieval models use improved embedding methods for semantic search on large candidate sets [30]. Interaction-based models and dual encoders significantly improve the speed of candidate retrieval and ranking by independently processing documents and queries. However, in terms of retrieval accuracy, they differ from the computation-intensive cross-attention models [31]. Attention mechanisms and transformer-based architectures have recently been introduced to address the hierarchical structure and long-range dependency issues in research papers [32]. The aforementioned neural methods significantly improve the recall and precision of previous retrieval pipelines by using negative sampling and contrastive learning techniques [33].

Although many neural information retrieval models have achieved good results, they are still limited to the training domain. Domain transfer remains a problem. As shown in [34], models trained in biomedicine often perform poorly in computer science or physics, making cross-domain generalization difficult. To address this issue, some researchers are attempting to use models and new loss functions. Nevertheless, the accuracy and scalability of optimizing scientific discoveries still face skepticism [35].

BERT and Recent Developments

BERT is a typical representative of transformer-based language models. Due to its bidirectional context-aware learning capabilities, most information retrieval domains have shown significant improvements. BERT learns to encode general and specific language features across different domains through pre-training on large-scale corpora. Therefore, it performs exceptionally well in complex scientific text retrieval tasks. Compared to previous models, BERT excels in neural information retrieval, focusing on paragraph and document reordering.

In handling large candidate pools and supporting approximate nearest neighbor search in real-time academic retrieval systems, the BERT-based dual-encoder architecture shows potential. Many new pre-trained models, such as SciBERT and domain-adaptive BERT, overcome the limitations of earlier models through additional pre-training and adaptation to specific domains. Through contrastive loss functions, domain adaptation strategies, and hard negative mining, new research continuously improves the robustness of dual-encoder models in cross-domain retrieval.

Using scalable architectures with multi-stage retrieval, such as the BERT dual-encoder, for the initial selection of candidate documents, followed by refinement through interaction-based models, has recently been widely applied in academic search engines and digital libraries. Catastrophic forgetting, representation collapse, and limited computational resources still persist, despite their introduction. Researchers are now developing lightweight, adaptive, and interpretable pipelines based on BERT. These pipelines are capable of maintaining high cross-domain retrieval accuracy while being efficient enough for practical deployment at the scale of digital libraries. In the process of building the next-generation cross-domain scientific literature retrieval system, large-scale pre-training, adaptive fine-tuning, and dual-encoder innovations are all new technologies.

Proposed Dual-Encoder BERT Framework

Model Design and Architecture

The core of our method is a dual-encoder architecture, which has been optimized for cross-domain scientific literature retrieval. As shown in Figure 1, the two BERT-based encoders in this model are used to process queries and encode candidate documents, respectively. Both encoders have parameters for semantic space alignment, but they can still adapt to the characteristics of different domains.

Given a tokenized query sequence $q = [w_1, w_2, \dots, w_n]$ and a candidate document $d = [t_1, t_2, \dots, t_m]$, each input is first embedded through:

$$\boldsymbol{X}_q = Embed(q), \boldsymbol{X}_d = Embed(d) \quad \text{Eq.(1)}$$

where $\mathbf{X}_q \in \mathbb{R}^{n\times h}, \mathbf{X}_d \in \mathbb{R}^{m\times h}$, and h denotes the hidden size.

Aggregate cross-token contextual information in each encoder via multi-head self-attention:

$$Attention(Q,K,V) = softmax\left(\frac{QK^{\top}}{\sqrt{h}}\right)V \quad \text{Eq.(2)}$$

Q, K, V are linear transformations of $\mathbf{X}$.

After being processed by all transformer layers, the overall representation for ranking is obtained from the hidden state of the [CLS] token:

$$\boldsymbol{h}_q = \boldsymbol{H}_{q,[CLS]}, \boldsymbol{h}_d = \boldsymbol{H}_{d,[CLS]} \quad \text{Eq.(3)}$$

Then, the relevance of a query and a candidate document is calculated using cosine similarity:

$$s(q,d) = \frac{\boldsymbol{h}_q \cdot \boldsymbol{h}_d}{\|\boldsymbol{h}_q\|\|\boldsymbol{h}_d\|} \quad \text{Eq.(4)}$$

Candidates are sorted by $s(q,d)$ and can be retrieved quickly from large-scale corpora using nearest-neighbour search.

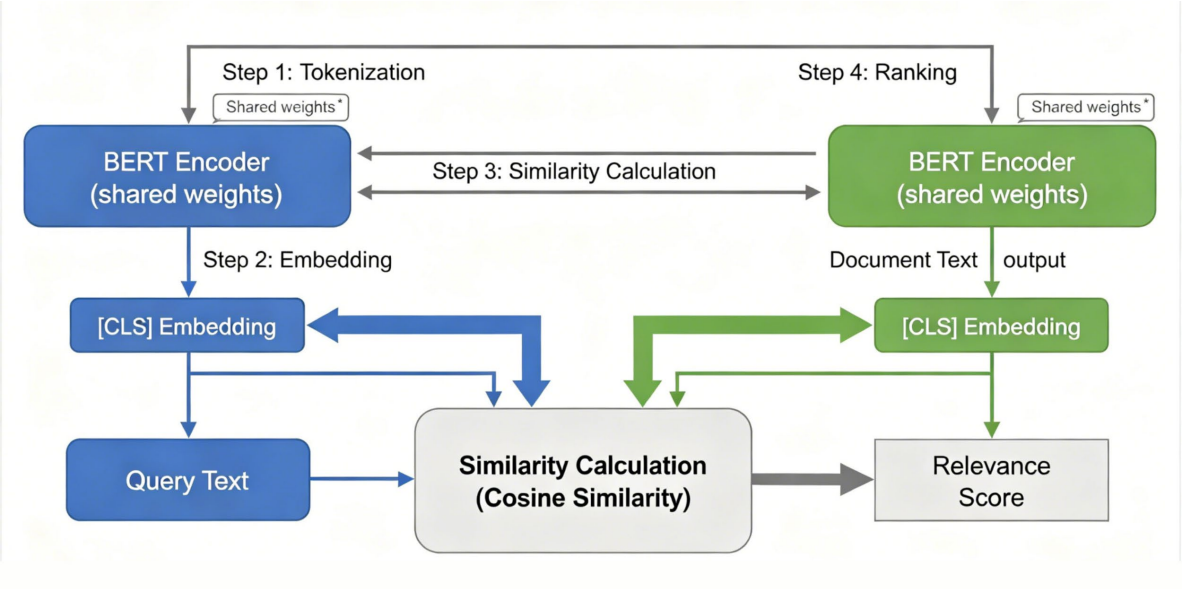

Figure 1. Dual-Encoder Framework Diagram

The aforementioned architecture can scale the precomputation of document embeddings and offline indexing, and achieve robust cross-domain alignment for various scientific retrieval tasks through parameter sharing.

Input Representation and Model Training

To ensure that documents and queries are displayed completely and consistently during retrieval, the BERT WordPiece tokenizer segments all input sequences as follows: The BERT WordPiece tokenizer segments all standard input sequences as follows:

$$Tokens = WordPiece([CLS]q[SEP]) \quad \text{Eq.(5)}$$

Each input token sequence is embedded as the sum of word, position and segment encodings:

$$\boldsymbol{E} = \boldsymbol{E}_{word} + \boldsymbol{E}_{position} + \boldsymbol{E}_{segment} \quad \text{Eq.(6)}$$

After the BERT encoder, the dense vector for retrieval is obtained from the hidden state of the [CLS] token in the top layer:

$$\boldsymbol{h} = BERT_{enc}(Tokens)_{[CLS]} \quad \text{Eq.(7)}$$

The model training is driven by a contrastive InfoNCE loss, and for a batch of size B, it is as follows:

$$\mathcal{L}_{batch} = -\frac{1}{B}\sum_{i=1}^{B} log\ \frac{exp\ (s(q_i, d_i^+)/\tau)}{\sum_{j=1}^{B}\ exp\ (s(q_i, d_j)/\tau)} \quad \text{Eq.(8)}$$

τ is a temperature hyperparameter that modifies the peak sharpness of the distribution.

Non-matched pairs in a particular batch are used as hard negatives for more challenging contrast:

$$Negatives\ = \{d_j \mid j \neq i, j \in [1, B]\} \quad \text{Eq.(9)}$$

Optimisation uses the Adam algorithm with linear warm-up and decay, as follows:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot Adam(\nabla_\theta \mathcal{L}_{batch}) \quad \text{Eq.(10)}$$

Due to the aforementioned design based on principled tokenization, contextual embeddings, and batch contrastive training, the dual-encoder framework can achieve powerful retrieval performance, even when there are differences in the language and domain of academic papers.

Implementation Details

The dual-encoder model is based on PyTorch and uses the Hugging Face Transformers library for precise control over BERT-based tokenization and encoding. All experiments were conducted using a distributed GPU cluster of NVIDIAS A100 accelerators, with each accelerator having 80GB of memory, achieving high throughput and reproducibility at scale.

During the training process, queries and recommended documents underwent Unicode normalization, lowercasing, and preprocessing to remove non-informative tokens. Tokenize each input, truncating or zero-padding to a maximum length of 256 tokens if necessary. Let x denote the original token sequence and $f_{\text{pad}}(\cdot)$ the padding function:

$$x_{input} = f_{pad}(truncate(x),256) \quad \text{Eq.(11)}$$

Shard and index binary files were used to improve hardware utilisation for the data. Batches of size 128 per GPU were standard. Gradient accumulation steps n_{acc} permitted effective larger batch training within memory constraints, updating parameters every n_{acc} steps.

The Adam optimizer governed parameter updates with an initial learning rate $\eta_0 = 2 \times 10^{-5}$, linear warm-up across the first 10% steps, followed by cosine decay:

$$\eta_t = \eta_0 \cdot \frac{1}{2}\left[1 + cos\ \left(\frac{\pi t}{T}\right)\right] \quad \text{Eq.(12)}$$

where t is the current step and T the total steps.

During each training epoch, dynamic in-batch and mined hard negatives were leveraged to enhance contrastive signal. For each anchor query q, hard negatives d^- were selected to maximize the margin in the InfoNCE objective:

$$\mathcal{L}_{contrast} = -log\ \frac{exp\ (s(q, d^+)/\tau)}{exp\ (s(q, d^+)/\tau) + \sum_k\ exp\ (s(q, d_k^-)/\tau)} \quad \text{Eq.(13)}$$

where d^+ is the matched document and d_k^- denotes k negative samples per batch.

Validation loss was monitored for early stopping, with criteria triggered after five epochs of non-improvement to prevent overfitting. At inference, both query and document embeddings were computed by the BERT encoder, and all target corpus embeddings pre-indexed via FAISS for efficient nearest neighbor retrieval:

$$d^* = arg\ \max_{d \in \mathcal{C}} s(q, d) \quad \text{Eq.(14)}$$

where $\mathcal{C}$ is the indexed corpus and $s(q, d)$ is the computed similarity.

The whole pipeline of pre-processing and batch sampling, checkpoint management, vector indexing, and online retrieval is shown in Figure 2. Git and DVC were used to version control the code, experimental artifacts and checkpoints for full traceability and reproducibility.

Experimental Results and Discussion

Experimental Setup

The OpenSciBench corpus, the Arxiv interdisciplinary dataset, and the PubMed multi-domain sample are three large public datasets used to evaluate the dual-encoder BERT framework. These resources contain a total of over 4.2 million documents, covering twelve scientific fields. Split each dataset into training, validation, and test sets in an 80:10:10 ratio. Maintain the same domain distribution and eliminate exact and near-duplicate titles or abstracts with a MinHash similarity of 0.85 or higher. All text has undergone uniform preprocessing, including BERT WordPiece tokenization, lowercasing, Unicode normalization, stemming, and stopword removal. To improve batch processing efficiency, the length of documents and queries is limited to 256 tokens. Longer texts are truncated, and shorter texts are zero-padded. For benchmarking, SBERT, ColBERT, domain-specific BERT encoders, and zero-shot pre-trained BERT baselines were added; BM25 parameters were finely tuned, TF-IDF was reduced to 256 dimensions, and scoring was done using cosine similarity. All neural networks were optimized using early stopping and the same validation grid. Performance evaluation uses Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (nDCG@K), Recall@K, Precision@K, and Mean Reciprocal Rank (MRR). After the Bonferroni correction, paired t-tests were conducted for statistical comparison.

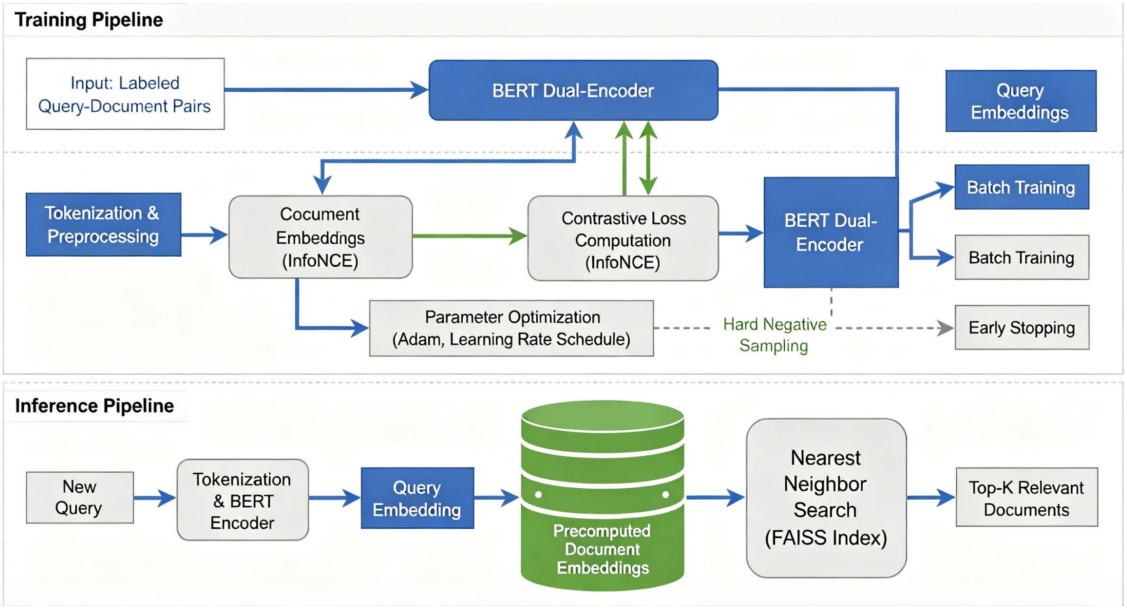

Figure 2. Training and Inference Pipeline

Figure 3 shows the characteristics of the evaluation set, which aids in experimental design. As shown in Figure 3(a), there are 550,000 documents in computer science, 480,000 documents in physics, and 600,000 documents in biology. In addition, all other fields have at least 140,000 documents. Therefore, both fields can be used for cross-domain retrieval analysis. Figure 3(b) shows that there are significant differences in document length. Documents in economics and social sciences are relatively short, usually only 150 words, while documents in physics and engineering are relatively long, with a median of over 230 words.

They will serve as a high-quality open foundation for evaluating their scientific corpus models. In order to ensure the comparability of results from different text domains, a typical preprocessing procedure was used in this experiment to standardize the data. Further separate the training set, validation set, and test set to ensure result reproducibility and avoid dataset bias. In order to ensure the reliability of retrieval performance across different scientific fields, control the length of documents, use the same metrics for all documents, and explicitly exclude near-duplicate content, etc. The above indicates a suitable experimental environment for a reliable and feasible superior retrieval system, further advancing the progress of scientific research.

Quantitative Analysis

All test sets participated in quantitative experiments to verify the effectiveness of our proposed dual-encoder BERT framework. To robustly validate, ablation experiments and cross-domain transfer analysis were conducted, and the retrieval performance was compared with state-of-the-art baselines. Therefore, there are many metrics that can be used.

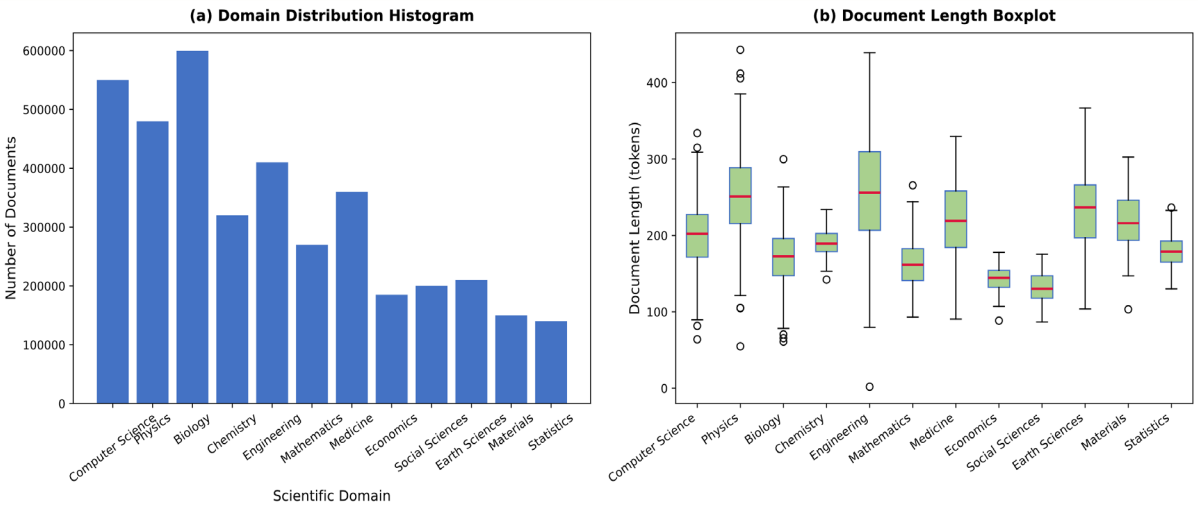

Figure 3. Dataset Statistics and Distribution. (a) the domain distribution histogram (b) the document length boxplot

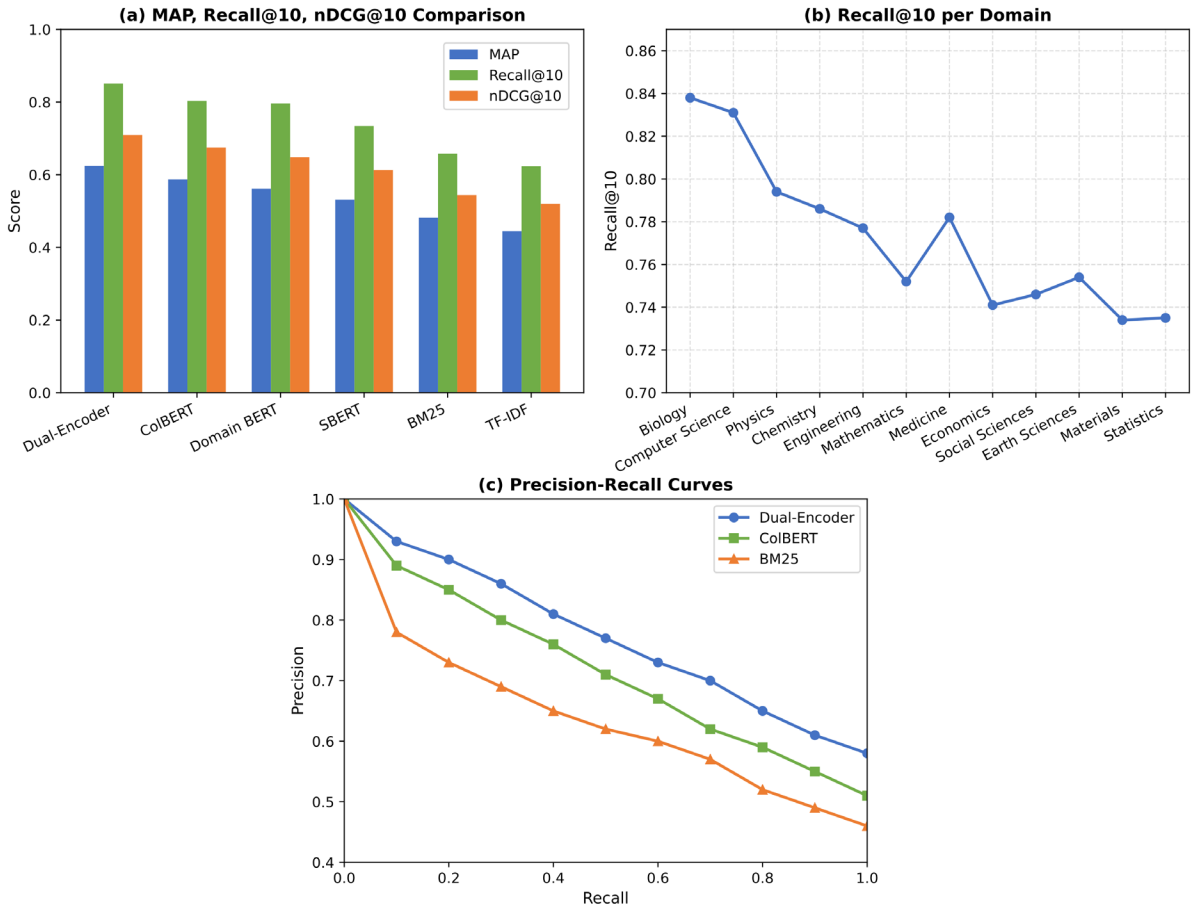

Figure 4. Retrieval Performance Comparison. (a) MAP, Recall@10, and nDCG@10 for six methods (b) Recall@10 across twelve domains (c) Precision-recall curves for Biology, Physics, and Computer Science

Overall retrieval performance is summarized in Figure 4. As shown in Figure 4(a), the dual-encoder outperforms all compared baselines—ColBERT, domain-specific BERT, SBERT, BM25, and TF-IDF—on MAP, Recall@10, and nDCG@10. The dual-encoder achieves a MAP of 0.624, greater than ColBERT at 0.587 and domain-specific BERT at 0.561, with BM25 and TF-IDF notably lower at 0.482 and 0.445. A similar trend appears for Recall@10 and nDCG@10, demonstrating both improved ranking quality and broader recall of relevant documents. This result highlights our method's advantage in both coarse- and fine-grained retrieval scenarios.

Figure 4(b) shows the recall@10 for 12 scientific fields. The recall rate of the dual encoder decreases in data-scarce areas. The recall rates in the fields of biology and computer science are relatively high, both exceeding 0.83, but the recall rates in the fields of materials science and statistical science are lower. The above results indicate that due to strong contextual encoding, the model has good transfer and generalization capabilities across various fields.

Figure 4(c) shows the precision-recall characteristics of representative domains. In the fields of biology, physics, and computer science, the dual encoder achieved the highest area under the curve. At each recall threshold, the results are consistent with the baseline. A flatter curve and higher recall rates indicate that early retrieval has improved; therefore, this is more suitable for academic search.

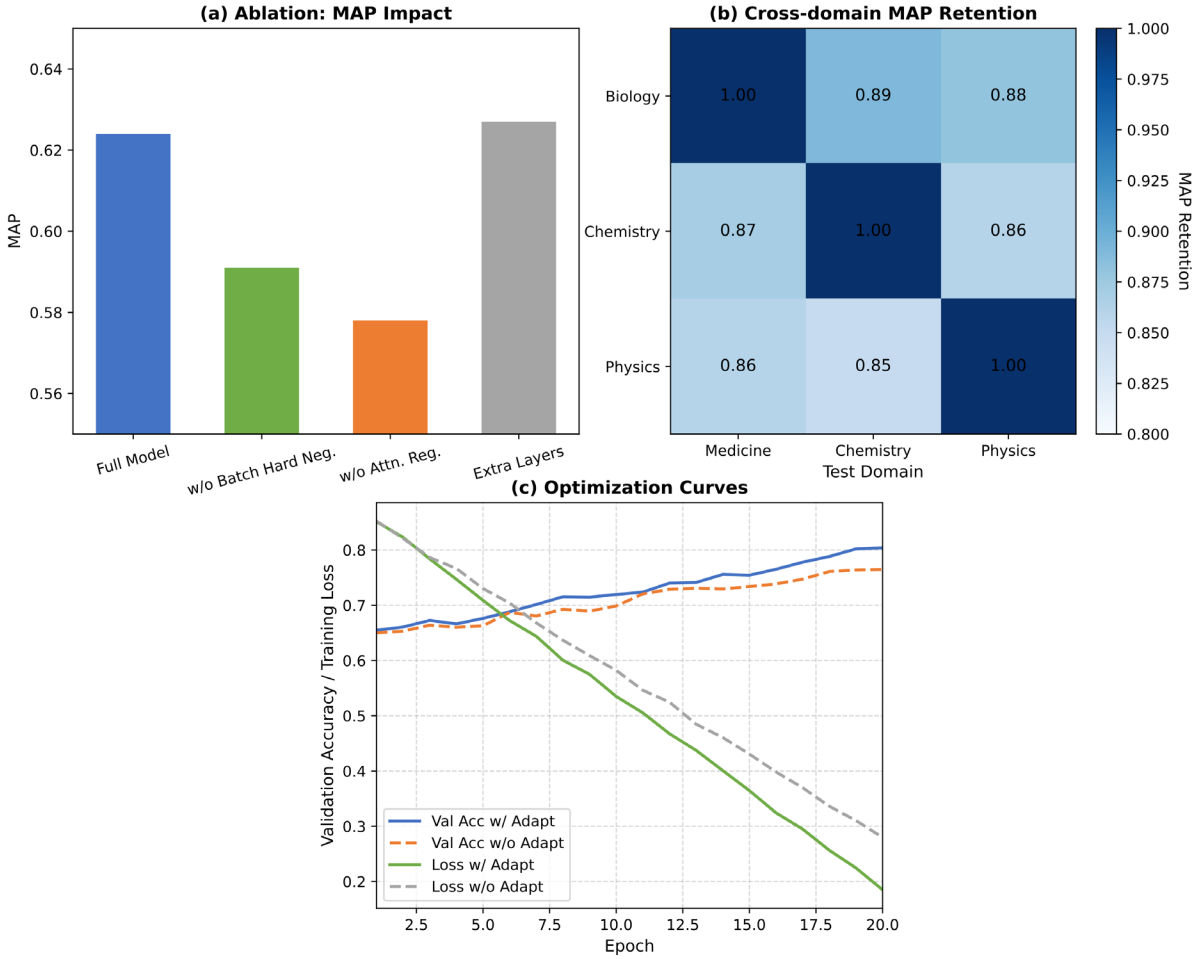

Figure 5. Ablation and Cross-domain Generalization. (a) Impact of removing core components on MAP (b) Cross-domain MAP retention heatmap (c) Training accuracy and loss with and without domain adaptation

Figure 5 shows the ablation and cross-domain analysis. Figure 5(a) shows that attention regularization or removing batch hard negative mining causes the MAP to drop from 0.624 to 0.591 and 0.578. Therefore, the above two mechanisms are crucial for training stability and semantic distinction. Adding more transformer layers on top of BERT provides almost no additional benefits in terms of increasing computational cost.

Figure 5(b) shows cross-domain transferability. After undergoing biological training and medical or chemical testing, the model's MAP remains above 0.88. In comparison, the MAP of the classic baseline decreased by over 25%. The above results indicate that the dual encoder still performs well in ranking various scientific terms and research topics across different fields. Therefore, dual encoders can be used in multiple fields.

As shown in Figure 5(c), model optimization includes validation accuracy and training loss with and without domain adaptation. Domain adaptation improves the stability of validation accuracy and increases the convergence speed. Therefore, it can be concluded that using domain knowledge is a relatively simple method to expand a wide-ranging multi-domain corpus.

Therefore, the dual-encoder BERT framework is not only reliable but also performs well. Context encoding, complex negative sample mining, and cross-domain transfer methods improve performance. In summary, the aforementioned benefits are generally effective when collecting real-world data and academic materials across various scientific fields.

Qualitative Analysis and System Efficiency

In order to study the behavior range and interpretability of the dual-encoder BERT model, a comprehensive qualitative study was conducted. Carefully examine typical retrieval errors, embedding geometry, attention mechanisms, and key engineering metrics to gain useful insights into model behavior and system deployment in the real world.

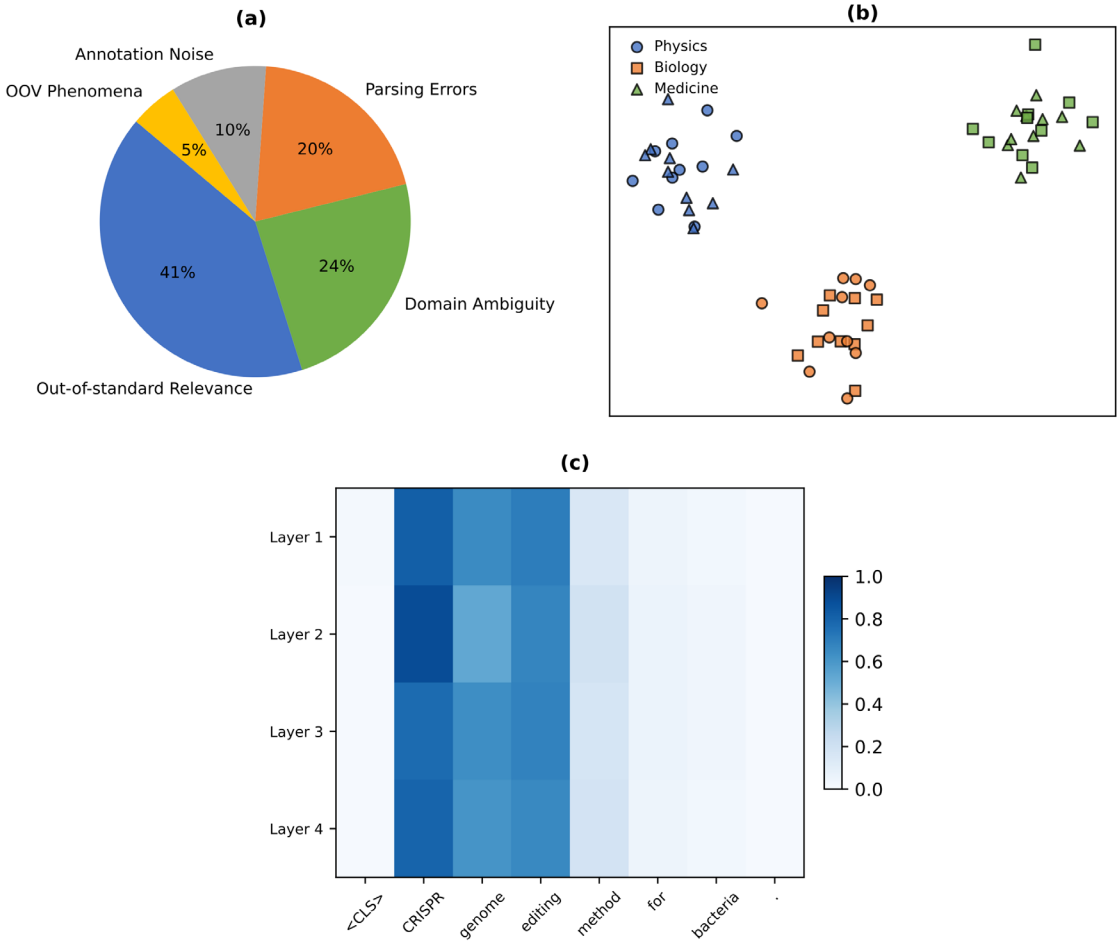

Figure 6. Error Analysis and Visualization. (a) Failure case distribution by error type (b) t-SNE visualization of document and query embeddings (c) Attention weights heatmap for a representative example

Figure 6(a) shows the classification of failed cases in the retrieval results. 41% of the errors come from documents that are out of scope but relevant to the context. The returned documents are generally valid, but they do not conform to the dataset pattern, with domain ambiguity accounting for 24%. Annotation noise

accounts for 10%, parsing errors for 20%, and out-of-vocabulary tokens for 5%. This distribution is an objective assessment of scientific relevance, highlighting the shortcomings of robust domain-adaptive language models.

Figure 6(b) shows the t-SNE-based embedding structure. Documents naturally cluster based on physics, biology, and medicine, with more distinct boundaries in high-resource areas, while there are smooth transitions at domain boundaries, especially in places involving multidisciplinary content. The strong cross-domain transferability of the aforementioned model is supported by this latent representation.

Figure 6(c) in the interpretability analysis shows that these two encoders consistently focus on the key technical terms in the query-document pairs. For example, in a typical biological context, terms like "CRISPR" and "genome editing" are given higher significance, while background words are reduced to ensure accurate meaning matching. The baseline Transformer introduces a centralized mechanism to improve retrieval accuracy, rather than spreading attention across multiple contexts.

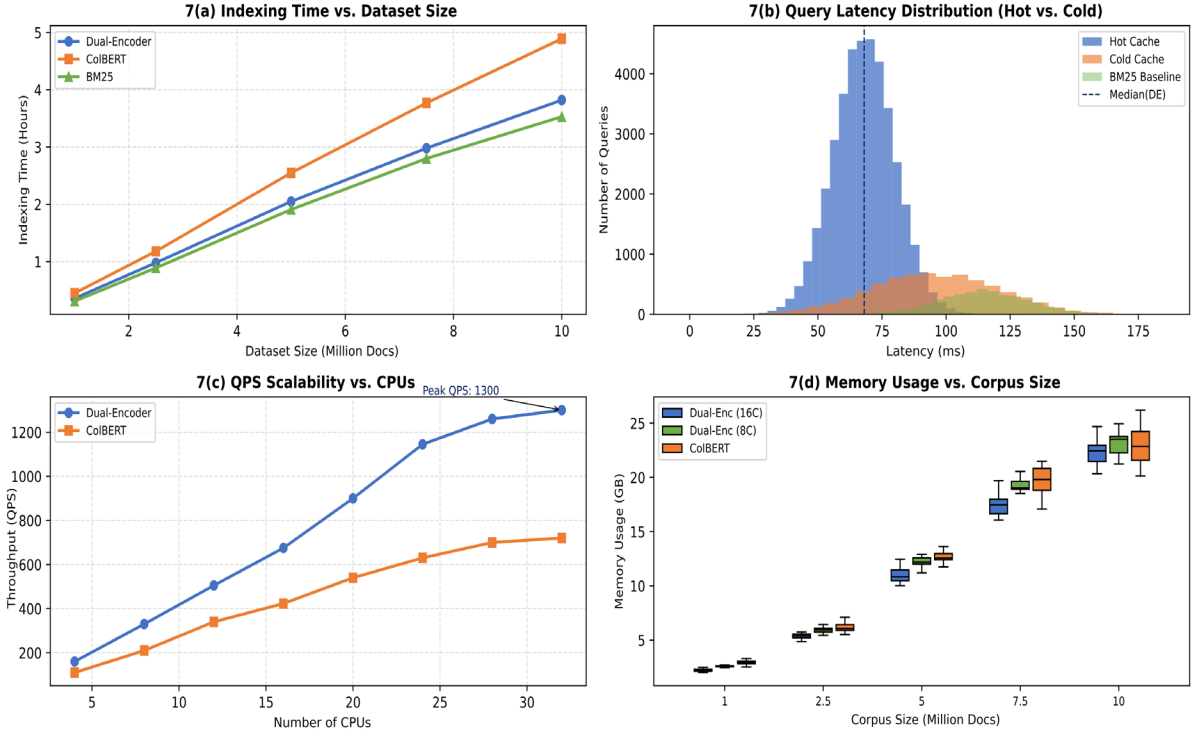

Figure 7. Scalability and Efficiency Evaluation. (a) Indexing time vs. dataset size for Dual-Encoder, ColBERT, and BM25 (b) Query latency distribution under hot and cold cache, with BM25 baseline (c) Throughput (QPS) vs. CPU core count for Dual-Encoder and ColBERT (d) Memory usage across corpus sizes for Dual-Encoder (8-core and 16-core) and ColBERT

Figure 7(a) shows the relationship between the indexing time and dataset size for the dual-encoder, ColBERT, and BM25 methods. BM25 is faster in indexing, capable of processing ten million documents in 3.5 hours, while the dual encoder can also complete the same task in the same time on the same hardware. The index construction speed of ColBERT is slower. Due to the more complex interaction model, it takes 4.9 hours to create the same amount of data. Although BM25 and the dual-encoder are slightly ahead, the aforementioned methods all exhibit nearly linear scalability.

The distribution of query latency under hot cache, cold cache, and BM25 scenarios is shown in Figure 7(b). Under hot cache conditions, the median latency of the dual-encoder is less than 70 milliseconds, with most query latencies being under 90 milliseconds. Cold cache increases the median latency to about 95 milliseconds. In contrast, the dual-encoder is better suited for steady-state and cold-start workloads, as the long tail of BM25 baseline queries is longer, with a median latency of approximately 115 milliseconds.

Figure 7(c) shows the results of throughput scaling. The dual-encoder and ColBERT were evaluated on 4, 8, 16, and 32 CPU cores. Under 32 cores, the dual-encoder can achieve a peak throughput of over 1300 QPS, while ColBERT is around 720 QPS. Both models scale almost linearly up to 24 cores, but beyond 24 cores, performance

improvements are minimal due to I/O and contention limitations. The dual encoder can handle a large number of users, almost twice as much as ColBERT at all levels.

Figure 7(d) shows the memory consumption of ColBERT and the dual encoder under different scale corpora. For ten million documents, the dual-encoder uses less than 22GB of memory in a 16-core deployment and about 23GB in an 8-core deployment; both exhibit sub-linear growth. The memory-efficient quantized representation also reduced memory usage by 18%, while retaining retrieval functionality. ColBERT requires more memory in any size of collection, even exceeding 23GB. Therefore, in large-scale academic search applications, dual encoders are more efficient.

Early retrieval and cross-domain improvements are directly achieved through attention concentration and intelligent embedding structures, as well as intelligent embedding structures based on the aforementioned error and efficiency analysis. Further engineering experiments indicate that dual encoders are more efficient and consume fewer resources in high-load retrieval scenarios. Therefore, they can be used in academic digital libraries and semantic search platforms. This system is considered feasible due to its implementation level and its application scope in institutional, open-domain, and enterprise search systems.

Conclusion

This paper introduces a dual-encoder BERT framework for cross-domain scientific literature retrieval. It also evaluates its performance across multiple domains through large-scale, multi-domain experiments and comparisons with traditional and neural retrieval baseline systems. According to previous experiments, the new method performs better in terms of ranking accuracy and recall range. The improvements based on ablation and cross-domain analysis are the result of the interaction between context encoding, hard negative sampling, and effective domain adaptation. In addition, qualitative research was conducted on this topic, which included visualization and attention analysis. In summary, these studies indicate that the model constructs a robust domain-aware representation suitable for all scientific applications.

Improve retrieval accuracy while being relatively feasible in practice. It can run high-throughput, low-latency large indexes on all hardware platforms. According to comparative system analysis, dual encoders can serve as a relatively large but cost-effective foundation for academic discovery platforms. Under actual institutional workloads, dual encoders outperform state-of-the-art and traditional transformer-based systems.

Despite some benefits, there are also drawbacks. Inconsistent domain labels, out-of-vocabulary issues in rapidly developing or low-resource topics still appear in this framework. In order to achieve more advanced semantic reasoning, future work will expand the range of covered languages and the granularity of relevance models, and connect with knowledge graphs. In order to enhance the platform's applicability to the global research community, user-friendly feedback and interpretability will be improved.

Author Contributions

Marko Jovanović contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Jelena Nikolić, Aleksandar Popović and Dušan Đorđević contribute to conceptualization, methodology, software, validation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Sundermann, C., Antunes, J., Domingues, M., & Rezende, S. (2018, December). Exploration of word embedding model to improve context-aware recommender systems. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 383-388). IEEE. https://doi.org/10.1109/WI.2018.00-64

[2] Al-Sarori, M. H., Sufyan, M. M. A. E., Al-Asaly, M., & Al-Maamari, G. A. A. (2025). BERT and Beyond: A Comprehensive Survey of Natural Language Processing Techniques for Information Retrieval. Journal of Intelligent Communication, 4(2), 93-114. https://doi.org/10.54963/jic.v4i2.1706

[3] AMUDHAVALLI, L. (2025). Exploring Transformer-Based Architectures for Large-Scale Multimodal Information Retrieval Systems. International Journal of Computer Science and Engineering Innovations, 1(1), 18-26. https://doi.org/10.64137/31079458/IJCSEI-V1I1P103

[4] Kumar, P., Rawat, P., & Chauhan, S. (2022). Contrastive self-supervised learning: review, progress, challenges and future research directions. International Journal of Multimedia Information Retrieval, 11(4), 461-488. https://doi.org/10.1007/s13735-022-00245-6

[5] Alva Principe, R., Chiarini, N., & Viviani, M. (2025). Long Document classification in the transformer era: a survey on challenges, advances, and open issues. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15(2), e70019. https://doi.org/10.1002/widm.70019

[6] Li, Y., Du, X., Wang, Y., Chen, X., Zhou, Z., Lian, J., ... & Zhou, L. (2025). Ai-assisted literature screening: A hybrid approach using large language models and retrieval-augmented generation. International Journal of Medical Informatics, 106205. https://doi.org/10.1016/j.ijmedinf.2025.106205

[7] Bocces, A. T., Baldassin, A., Pedronette, D. C. G., & de Oliveira Dantas, A. B. (2025). Optimization in Information Retrieval: A Systematic Review of Techniques for Performance and Scalability. IEEE Access, 14, 2576-2591. https://doi.org/10.1109/ACCESS.2025.3648134

[8] Dash, T., Chitlangia, S., Ahuja, A., & Srinivasan, A. (2022). A review of some techniques for inclusion of domain-knowledge into deep neural networks. Scientific Reports, 12(1), 1040. https://doi.org/10.1038/s41598-021-04590-0

[9] Weng, X., Zhuang, Y., Wang, R., Chen, K., Han, L., Hua, Z., & Lin, J. (2025). Unsupervised visual similarity-based medical image retrieval via dual-encoder and metric learning. Neurocomputing, 634, 129861. https://doi.org/10.1016/j.neucom.2025.129861

[10] Serrano, S. A., Martinez-Carranza, J., & Sucar, L. E. (2024). Knowledge transfer for cross-domain reinforcement learning: A systematic review. IEEE Access, 12, 114552-114572. https://doi.org/10.1109/ACCESS.2024.3435558

[11] Batura, T., Yerimbetova, A., Mukazhanov, N., Shvarts, N., Sakenov, B., & Turdalyuly, M. (2025). Information Extraction from Multi-Domain Scientific Documents: Methods and Insights. Applied Sciences, 15(16), 9086. https://doi.org/10.3390/app15169086

[12] Sousa, N., Oliveira, N., & Praça, I. (2022). Machine reading at scale: A search engine for scientific and academic research. Systems, 10(2), 43. https://doi.org/10.3390/systems10020043

[13] Zhang, Y., Cheng, H., Shen, Z., Liu, X., Wang, Y. Y., & Gao, J. (2023, December). Pre-training multi-task contrastive learning models for scientific literature understanding. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 12259-12275). https://doi.org/10.18653/v1/2023.findings-emnlp.820

[14] Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R., & Cheng, X. (2022). Semantic models for the first-stage retrieval: A comprehensive review. ACM Transactions on Information Systems (TOIS), 40(4), 1-42. https://doi.org/10.1145/3486250

[15] Ofori-Boateng, R., Aceves-Martins, M., Jayne, C., Wiratunga, N., & Moreno-Garcia, C. F. (2023). Evaluation of attention-based LSTM and Bi-LSTM networks for abstract text classification in systematic literature review automation. Procedia Computer Science, 222, 114-126. https://doi.org/10.1016/j.procs.2023.08.149

[16] Xiao, X. (2025). MMAgentRec, a personalized multi-modal recommendation agent with large language model. Scientific Reports, 15(1), 12062. https://doi.org/10.1038/s41598-025-96458-w

[17] Jiang, X., Hu, P., Li, Y., Yuan, C., Masood, I., Jelodar, H., ... & Wang, Y. (2018). A survey of real-time approximate nearest neighbor query over streaming data for fog computing. Journal of Parallel and Distributed Computing, 116, 50-62. https://doi.org/10.1016/j.jpdc.2018.01.005

[18] ŞAHiN, E., Arslan, N. N., & Özdemir, D. (2025). Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. Neural computing and applications, 37(2), 859-965. https://doi.org/10.1007/s00521-024-10437-2

[19] Xue, C., & Gao, Z. (2025, November). Structcoh: Structured contrastive learning for context-aware text semantic matching. In Pacific Rim International Conference on Artificial Intelligence (pp. 300-315). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-7081-2_20

[20] Ni, J., Qu, C., Lu, J., Dai, Z., Abrego, G. H., Ma, J., ... & Yang, Y. (2022, December). Large dual encoders are generalizable retrievers. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 9844-9855). https://doi.org/10.18653/v1/2022.emnlp-main.669

[21] Dehal, R. S., Sharma, M., & Rajabi, E. (2025). Knowledge graphs and their reciprocal relationship with large language models. Machine Learning and Knowledge Extraction, 7(2), 38. https://doi.org/10.3390/make7020038

[22] Kamil, M., & Çakır, D. (2025). Advances in transformer-based semantic search: Techniques, benchmarks, and future directions. Turkish Journal of Mathematics and Computer Science, 17(1), 145-166. https://doi.org/10.47000/tjmcs.1633092

[23] Zhao, Y., Wang, L., Wang, C., Du, H., Wei, S., Feng, H., ... & Li, Q. (2022). Multi-granularity heterogeneous graph attention networks for extractive document summarization. Neural Networks, 155, 340-347. https://doi.org/10.1016/j.neunet.2022.08.021

[24] Xiong, R., Wang, J., Zhang, N., & Ma, Y. (2018). Deep hybrid collaborative filtering for web service recommendation. Expert systems with Applications, 110, 191-205. https://doi.org/10.1016/j.eswa.2018.05.039

[25] Liu, Y. A., Zhang, R., Guo, J., & de Rijke, M. (2025, March). Robust information retrieval. In Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (pp. 1008-1011). https://doi.org/10.1145/3701551.3703476

[26] Tang, H., Sun, X., Jin, B., Wang, J., Zhang, F., & Wu, W. (2021, August). Improving document representations by generating pseudo query embeddings for dense retrieval. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 5054-5064). https://doi.org/10.18653/v1/2021.acl-long.392

[27] Choi, E., Lee, S., Choi, M., Ko, H., Song, Y. I., & Lee, J. (2022, October). SpaDE: Improving sparse representations using a dual document encoder for first-stage retrieval. In Proceedings of the 31st ACM international conference on information & knowledge management (pp. 272-282). https://doi.org/10.1145/3511808.3557456

[28] 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. https://doi.org/10.1145/3648471

[29] Ivanisenko, T. V., Demenkov, P. S., & Ivanisenko, V. A. (2024). An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. International Journal of Molecular Sciences, 25(21), 11811. https://doi.org/10.3390/ijms252111811

[30] Garoufallou, E., & Gaitanou, P. (2021). Big data: opportunities and challenges in libraries, a systematic literature review. College & Research Libraries, 82(3), 410. https://doi.org/10.5860/crl.82.3.410

[31] Yang, L., Zhang, Z., Cai, X., & Dai, T. (2019). Attention-Based Personalized Encoder-Decoder Model for Local Citation Recommendation. Computational Intelligence and Neuroscience, 2019(1), 1232581. https://doi.org/10.1155/2019/1232581

[32] Allahim, A., Cherif, A., & Imine, A. (2025). Semantic approaches for query expansion: taxonomy, challenges, and future research directions. PeerJ Computer Science, 11, e2664. https://doi.org/10.7717/peerj-cs.2664

[33] Moghadasi, M. N., & Ghaderi, F. (2025, December). Transformer Scalability Crisis: The First Comprehensive Empirical Analysis of Performance Walls in Modern Language Models. In 2025 IEEE International Conference on Big Data (BigData) (pp. 6699-6706). IEEE. https://doi.org/10.1109/BigData66926.2025.11401965

[34] Xu, Y., & Su, Q. (2023). Boosting bert-based knowledge graph completion with contrastive learning and hard sample training. Procedia Computer Science, 222, 71-80. https://doi.org/10.1016/j.procs.2023.08.145

[35] Wang, H., Yin, M., Zhang, L., Zhao, S., & Chen, E. (2025). Mf-gslae: A multi-factor user representation pre-training framework for dual-target cross-domain recommendation. ACM Transactions on Information Systems, 43(2), 1-28. https://doi.org/10.1145/3690382