Hybrid Learning-to-Rank Approach for Complex Information Retrieval Systems

Application to Biomedical Question Answering

Fatma Zohra Bessai-Mechmache[1], Yasmine Hanifi[2], Damia Lyna Ait Idir[3]
CERIST Research Center, Algiers, Algeria[1]
Computer Science Department, University of Algiers, Algiers, Algeria[2, 3]

***Abstract*—Biomedical question answering presents significant challenges due to the complexity of biomedical language and the need for precise information retrieval. This study aims to improve the performance of a biomedical information retrieval system through a hybrid learning-to-rank framework. Specifically, we combine lexical (BM25) and semantic (BioBERT) representations to form hybrid inputs for RankFormer, a transformer-based ranking model. This hybrid representation captures both surface-level term matching and deep contextual understanding. Experiments conducted on the BioASQ dataset show that our approach achieves better ranking performance compared to the standalone lexical or neural baselines, reaching a MAP@10 of 0.9614 and an nDCG@10 of 0.9320. These results highlight the effectiveness of hybrid input representations in enhancing biomedical answer ranking.**

***Keywords*—Learning-to-rank; information retrieval; hybrid learning-to-rank; transformer-based ranking model; biomedical question answering**

I. Introduction

Ranking in large-scale information retrieval can be seen as a multi-stage process where efficiency, relevance, and ranking stability must be jointly addressed.

Biomedical question answering (BQA) sits at the intersection of natural language processing and information retrieval (IR), aiming to return relevant answers to complex biomedical questions. The volume of biomedical literature continues to grow exponentially, and the terminology itself is often highly specialized and ambiguous. Together, these factors tend to make effective information retrieval genuinely challenging in real-world biomedical settings.

Traditional lexical retrieval models, with BM25 as the most common example, rely on exact term matching and have, for years, served as solid baselines in information retrieval. They are fast, relatively simple to deploy, and generally do a good job when surface-level word overlap is enough. However, such models tend to struggle with semantic variation, including synonymy and context-dependent meanings, all of which appear frequently in biomedical writing.

Neural language models like BioBERT take a different direction by using contextualized representations that aim to capture semantic relationships beyond lexical similarity. Although these models have demonstrated strong performance in biomedical natural language understanding tasks, they can underperform when precise term matching matters. They typically require computationally intensive re-ranking.

Recent studies suggest that mixing lexical and semantic signals can alleviate the limitations of individual models. Hybrid retrieval strategies that use sparse (exact relevance signal) with dense (contextual relevance signal) representations often perform better across multiple benchmarks, and the gains tend to be more noticeable in specialized domains. In the biomedical domain, where both precise terminology and deep contextual understanding are crucial, such strategies are particularly relevant.

In this work, we aim to improve the performance of biomedical information retrieval systems by adopting a hybrid learning-to-rank approach. Specifically, we propose a method that brings together lexical scores from BM25 and contextual embeddings derived from BioBERT to form a hybrid input for RankFormer, a transformer-based listwise ranking model. Rather than treating sparse and dense relevance signals separately, the model is trained to learn from them jointly within a supervised setting, which appears to encourage rankings that are not only more accurate but also more sensitive to meaning and context. We experiment with the proposed approach on the BioASQ Task B 2024 dataset, where the results demonstrate that combining hybrid representations with listwise learning can lead to noticeable gains over both purely lexical methods and standalone neural baselines.

II. Related Work

Information retrieval is the process of locating information items that best align with a user's query [1], [2]. In biomedical contexts, this task becomes especially demanding. The huge volume of biomedical literature, along with its domain-specific terminology, means that retrieval systems often play a practical role in clinical decision support and day-to-day biomedical research [3]. Not surprisingly, a wide range of methods has been proposed over time to improve retrieval quality, from classic lexical techniques to more recent neural models and, increasingly, hybrid approaches that try to balance precision with semantic understanding.

A. Lexical Methods (Sparse Retrieval Methods)

Traditionally, lexical approaches relying on word overlap have been the foundation of information retrieval [4]. Sparse retrieval methods use inverted indexes to link terms to documents, which helps explain why they remain

computationally efficient even at a large scale. Over the years, researchers have proposed a range of weighting and normalization schemes, rising to multiple Term Frequency-Inverse Document Frequency (TF-IDF) models [5]. Because it offers a reasonable balance between effectiveness and simplicity, BM25 [6] is one of the sparse models that is often viewed as the default choice in real-world systems. In practice, it processes documents as bags of words and assigns scores based on the presence of query terms independently of their position or proximity in the text [7].

A key strength of BM25 lies in its ability to normalize document length and address query term saturation. This enables it to avoid bias toward either very short to very long documents and to reduce the disparate influence of excessively repeated terms. BM25 has been extended to incorporate multiple document fields and specify how their individual scores should be combined [8].

Lexical models continue to be popular largely because they are fast and easy to deploy at scale. In many general-purpose retrieval settings, lexical methods are usually enough. However, when the language becomes more specialized, as in biomedical texts, synonymy and contextual ambiguity are common, and methods based solely on word overlap are likely to miss relevant documents even when the underlying concepts align [9].

B. Semantic Methods (Dense Retrieval Methods)

Efforts to improve information retrieval systems have increasingly turned toward neural-based semantic retrieval approaches. Early work in this category relied on convolutional and recurrent neural networks to represent textual data. In practice, both queries and documents are mapped into a shared vector space assumed to capture more nuanced representations of meaning. Relevance is then estimated through similarity functions trained to reflect semantic alignment between these representations [10], [11].

Additionally, the neural models benefit from their ability to learn non-linear text representations, which explain why they often outperform traditional lexical approaches [11]. A more noticeable advancement in the field came with the introduction of transformer models. Transformer-based models such as BERT [12] and GPT [13], thanks to the attention mechanism, have significantly improved retrieval performance by enabling the model to focus attention on the most relevant parts of both queries and documents during the matching process [14].

Neural models such as BioBERT were introduced to move beyond surface-level term matching and toward a richer notion of semantic similarity. As a domain-adapted version of BERT, BioBERT is pre-trained on large-scale biomedical texts. This model has reported strong performance on biomedical question answering and document ranking benchmarks [15]. Neural models offer an improved ability to relate concepts that are not lexically identical. For BQA systems, where relevant evidence is often paraphrased rather than repeated verbatim, this contextual sensitivity is crucial [16].

C. Hybrid Methods

The development of hybrid approaches is motivated by the strengths and complementary results produced by sparse and dense retrieval models [1]. Hybrid methods typically merge lexical and semantic scores using fusion techniques. Among the most widely used strategies are Convex Combination and Reciprocal Rank Fusion (RRF).

Although both strategies are widely used, Convex Combination appears to be the more flexible option in many settings. It requires only minimal score normalization and, according to prior studies, has in several cases outperformed RRF [17]. One practical advantage, however, is that it tends to be sample-efficient, adapting reasonably well to specific domains with relatively few examples.

A common strategy involves using BM25 to retrieve an initial set of candidate documents, which are then re-ranked using a dense encoder such as BioBERT. This two-stage architecture has been shown to significantly improve retrieval performance in various QA benchmarks. More recent work has pushed this integration further by directly fusing lexical and dense scores.

D. Learning-to-Rank

Learning-to-rank (LTR) is designed as a supervised learning approach for ordering documents by how relevant they are to a given query. LTR has become an important component of modern information retrieval systems, especially in settings where simple heuristics no longer suffice. That said, LTR actually covers a range of modeling choices rather than a single, unified method.

Most LTR techniques are commonly organized into three broad paradigms [18], each of which frames the ranking problem in a slightly different way. The pointwise approach, for instance, treats the ranking problem as a classification task, where each document is assigned a relevance score independently of the others. This setup is attractive because it is easy to implement and fits neatly within existing machine learning pipelines. However, by scoring documents independently, it arguably misses a key aspect of ranking because relevance is often relative and emerges from direct competition among documents shown for the same query.

Pairwise methods try to close this gap by shifting the learning objective toward preferences between pairs of documents. Instead of asking whether a document is relevant in absolute terms, the model learns which of two documents should be ranked higher. RankNet, introduced by Burges et al. [18], is a well-known example; it uses a neural network to estimate the probability that one document should be ranked ahead of another.

Listwise methods take a more holistic view by considering the entire set of documents associated with a query during training. One of the more influential examples is ListNet [19]. ListNet defines a probability distribution over possible permutations of the result list. This formulation appears to let the model capture interactions among documents more directly.

Recent progress in deep learning has pushed learning-to-rank models toward transformer-based approaches, largely due to their ability to capture complex contextual relationships. One such transformer-based listwise LTR model that incorporates traditional listwise objectives with innovative listwide assessment objectives is RankFormer [20]. This design appears

to help the model capture dependencies across an entire document list, leading to improved ranking performance.

Our work follows this line of research but takes a different angle. We propose a hybrid approach in which both BM25 scores and BioBERT embeddings are provided as input signals to a RankFormer-based re-ranker. Rather than treating lexical retrieval and semantic re-ranking as two clearly separated stages, we integrate sparse and dense features from the outset. The idea here is that BioBERT captures domain-specific semantics, whereas BM25 tends to anchor the ranking in exact term matches. In order to take benefit of the complementary advantages of both sparse and dense retrieval for more precise biomedical passage ranking, the two are combined early in the ranking process.

III. Methodology

The proposed methodology adopts a hierarchical learning-to-rank framework to progressively enhance document relevance for biomedical question answering. To take advantage of complementary relevance signals at different stages of the ranking process, the approach combines supervised listwise learning with semantic re-ranking and lexical retrieval. Three sequential steps are used to demonstrate the workflow: 1) BM25 is used for initial candidate retrieval; 2) BioBERT is used for semantic re-ranking; and 3) a hybrid RankFormer model is used for final ranking.

This multi-stage strategy seeks to balance efficiency and effectiveness by leveraging the strong recall of sparse retrieval with the semantic sensitivity of dense representations and the optimization capabilities of learning-to-rank models.

For a given query, we start by keeping the top 50 documents returned by BM25. At this stage, the goal is coverage rather than finesse. BioBERT then re-ranks this shortlist, filtering it down to the 25 documents that appear to align most closely with the query at a semantic level. These candidates are then passed to RankFormer, which generates the final top-10 ranking. This multi-stage ranking strategy has proven to be effective in advanced information retrieval systems, as it brings together the efficiency of sparse retrieval with the precision of dense methods and learning-to-rank techniques [21].

A. Dataset

We carry out our experiments on the BioASQ Task B 2024 dataset, a reference benchmark for biomedical question answering. The dataset consists of expert-formulated biomedical questions. Each question is linked to a set of supporting materials: relevant PubMed documents, text snippets, biomedical concepts, and expert answers curated by domain specialists [22].

BioASQ Task B queries are natural-language biomedical questions, categorized into four classes: Factoid questions (single entity answers), List questions (multiple entity answers), Yes/No questions, and Summary (ideal answer) questions.

For retrieval and ranking purposes, the dataset is preprocessed into three components:

- Queries: Question texts are extracted from the original JSON files and stored in JSONL format, where each entry contains a unique query identifier and the query body.
- Documents: Document collections are constructed by concatenating snippets associated with the same source URL, resulting in a dense textual corpus indexed using the Pyserini [23] toolkit.
- Relevance Judgments (Qrels): Ground-truth relevance annotations are formatted according to the standard TREC qrels specification, linking each query to relevant documents.

This preprocessing task ensures compatibility with both sparse and dense retrieval models and provides a reliable evaluation framework for learning-to-rank experiments.

B. Initial Lexical Retrieval with BM25

The suggested methodology starts with BM25, a probabilistic sparse retrieval model that is frequently used in information retrieval because of its interpretability and robustness. BM25 ranks documents based on exact query term matching using term frequency saturation and document length normalization.

We test four distinct BM25 configurations in order to examine the effects of BM25 hyperparameters. Precisely, we vary the term frequency parameter "k_1" and the length normalization parameter "b":

- $k_1 = 0.9$, $b = 0.4$
- $k_1 = 0.9$, $b = 0.8$
- $k_1 = 1.2$, $b = 0.4$
- $k_1 = 1.2$, $b = 0.8$

For each query, the top 50 documents are retrieved from the indexed corpus. This candidate set is intentionally kept broad in order to maximize recall, ensuring that potentially relevant documents are preserved for subsequent re-ranking stages.

C. Semantic Re-ranking with BioBERT

In the second stage, we move beyond lexical matching and introduce semantic re-ranking with BioBERT, a transformer model pre-trained on large-scale biomedical datasets. The idea is that once BM25 has gathered a set of plausible candidates, BioBERT reconsiders them through a contextual goal. Instead of relying purely on term overlap, it can account for synonymy and the kinds of domain-specific expressions that often appear in biomedical writing.

At first, we applied BioBERT in its pre-trained form, using it to score the top 50 documents retrieved by BM25 for each query. Nevertheless, initial tests show that task sensitivity is low in this zero-shot setting. The model is fine-tuned on the BioASQ relevance judgments in order to overcome this constraint.

Fine-tuning data are constructed by pairing queries with candidate documents and assigning binary relevance labels derived from the qrels. A subset of 1,000 query-document pairs is randomly sampled and split into 80% for training and 20% for validation. Inputs are tokenized using the BioBERT tokenizer,

with truncation and padding applied to a maximum sequence length of 512 tokens.

The model is fine-tuned as a binary classifier for three epochs using a learning rate of 2e-5, a batch size of 8, and a weight decay of 0.01. Validation accuracy is used to monitor training progress. After fine-tuning, BioBERT demonstrates improved relevance prediction and is used to re-rank the BM25 candidate set. The top 25 documents per query are retained for the final ranking stage. This semantic re-ranking stage serves as a critical bridge between traditional lexical matching and transformer-based contextual understanding, yielding more precise retrieval results in the biomedical domain.

D. RankFormer Hybrid Learning-to-Rank Model

The final ranking stage employs RankFormer, a transformer-based listwise learning-to-rank model capable of modeling inter-document dependencies within a candidate list. Unlike conventional pipelines that treat lexical and semantic re-ranking as isolated steps, our approach integrates hybrid relevance signals directly into the RankFormer training process.

Each training instance is defined as a triplet (q, D, L), where q denotes a query, D = {d_1, d_2, ..., d_n} is the set of candidate documents, and L represents a list-level supervision signal referred to as the listwise label. Each document d_i is associated with a document ID, a textual representation, and a relevance label {0, 1}, a predicted score, and a rank reflecting its position within the list.

To generate input scores for RankFormer, we compute a normalized linear combination of BM25 scores and fine-tuned BioBERT scores. The hybrid score for a document is given by Eq. (1) [24]:

$$score = \alpha \cdot scorebm25 + (1 - \alpha) \cdot scorebiobert \quad (1)$$

where, $\alpha \in [0, 1]$ controls the relative contribution of lexical and semantic signals. In our experiments, α is empirically set to 0.5. The choice of setting α=0.5 is driven by the objective of achieving a balanced integration of heterogeneous scoring signals. In biomedical document retrieval, BM25 ensures precise lexical matching for highly specialized terminology, whereas BioBERT captures richer semantic context and synonymy. Assigning equal weights prevents the RankFormer input from being biased toward either surface-level keyword matching or purely latent representations, thereby maintaining the precision of the former while benefiting from the recall of the latter.

In addition to document-level relevance labels, RankFormer incorporates list-level supervision. A binary listwide label "L" is assigned to each query: the label is set to 1 if the candidate list contains at least three relevant documents, and 0 otherwise. This supervision enables the model to assess global ranking quality beyond individual document relevance.

To optimize the model, the training objective combines a listwise ranking loss and a listwide loss.

The listwise loss encourages correct relative ordering of documents within each query candidate list. It follows a Softmax-based ranking loss [19], inspired by ListNet and defined by Eq. (2) as follows:

$$Llistwise = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{|D_i|} (\log \sum_{k=j}^{|D_i|} \exp(s_{i,k}) - s_{i,j}) \quad (2)$$

where, S_{ij} denotes the predicted score for the j-th document in the i-th list and N is the total number of training queries.

The listwide loss is defined by Eq. (3) as a mean squared error (MSE) [25] between the predicted list quality and the true listwide label.

$$Llistwide = \frac{1}{N} \sum_{i=1}^{N} (l_i - \widehat{l_i})^2 \quad (3)$$

where, l_i is the true listwide label for the i-th query, and $\widehat{l_i}$ is the predicted list quality output by the model.

As specified in Eq. (4), the final training objective function is expressed as a weighted combination of both losses.

$$Ltotal = \alpha \cdot Llistwise + (1 - \alpha) \cdot Llistwide \quad (4)$$

where, α is set to 0.7, to prioritize accurate document ranking while maintaining sensitivity to list-level performance.

This hybrid supervision strategy allows RankFormer to jointly optimize fine-grained relevance ordering and holistic ranking performance, making it particularly suitable for biomedical information retrieval tasks.

The RankFormer-Hybrid model architecture is built upon a BioBERT backbone, which serves as a pre-trained encoder to extract domain-specific contextual embeddings. To manage overhead, these high-dimensional representations are projected into a 128-dimensional space via a linear transformation layer. The primary ranking operations are then conducted by a streamlined Transformer encoder consisting of two layers and four attention heads. Within this framework, query-document fusion is achieved by appending the query representation to each document vector, allowing the model to capture fine-grained, query-dependent relevance signals. The architecture concludes with a dual-level scoring mechanism, an individual linear layer computes local scores for each document, while a separate listwide score is generated by averaging document representations through a global linear layer.

As illustrated in Table I, the entire architecture is optimized using a weighted hybrid loss function, Eq. (4), which balances the precision of relative document ordering with the overall coherence of the retrieved list.

TABLE I. RANKFORMER-HYBRID MODEL ARCHITECTURE

Component	Description
Pre-trained encoder	BioBERT (dmis-lab/biobert-base-cased-v1.1)
Linear projection layer	Linear to a 128-dimensional space
Transformer encoder	2 layers, 4 attention heads, batch-first = true
Query-document fusion	Adding the query representation to each document
Score estimation	Individual linear layer for each document (score-layer)
Listwide score	Average of documents → linear layer (listwide-layer)
Learning loss	0.7· Loss-listwise + 0.3· Loss-listwide

E. Evaluation Metrics

Ranking performance is evaluated using standard information retrieval metrics at a cutoff of k = 10, consistent with the BioASQ Task B setting. The metrics include MAP@10, MRR@10, nDCG@10, Precision@10, and Recall@10. Together, these metrics provide a comprehensive assessment of ranking accuracy, early precision, and retrieval completeness.

A brief definition of these evaluation metrics [26] is as follows:

- Mean Average Precision at 10 (MAP@10) computes the average precision across all relevant documents within the top 10 retrieved, weighted by their position. This metric emphasizes both precision and the ranking order of relevant documents.
- Mean Reciprocal Rank at 10 (MRR@10) captures the position of the first relevant document in the ranking. It is defined as the reciprocal of the rank at which the first relevant document appears, providing a measure of how quickly a system can retrieve a relevant result.
- Normalized Discounted Cumulative Gain at 10 (nDCG@10) evaluates the quality of the ranking by assigning higher importance to relevant documents appearing higher in the list. It accounts for graded relevance and uses a logarithm discount factor to penalize lower-ranked relevant documents.
- Precision at 10 (P@10) measures the proportion of relevant documents among the top 10 retrieved items, offering an intuitive assessment of immediate result relevance.

Recall at 10 (R@10) quantifies the fraction of all relevant documents that are retrieved within the top 10 results, thereby reflecting the model's overall retrieval coverage.

These complementary metrics provide a comprehensive evaluation of the studied models, capturing different aspects of retrieval performance, including completeness, early precision, and ranking accuracy. Following the proposed methodology, the next sections present and analyze the experimental results obtained within this evaluation framework.

IV. Experimental Results and Discussion

The experimental results demonstrate a clear and consistent improvement in retrieval performance as the ranking pipeline evolves from lexical to dense and finally to supervised learning-to-rank approaches. This progression highlights the complementary nature of sparse and dense relevance signals and confirms the central role of supervised ranking in biomedical information retrieval.

A. BM25 Results

The BM25 baseline provides a strong lexical foundation, achieving competitive precision-oriented metrics when properly tuned. These results are consistent with prior findings showing that sparse retrieval models remain effective in biomedical settings when domain-specific terminology is well represented. However, the observed limitations in recall and nDCG indicate that lexical matching alone is insufficient to fully capture the semantic variability inherent in biomedical queries and documents.

Table II presents the evaluation results for the four configurations of the parameters k_1 and b across the five standard retrieval metrics at rank 10.

TABLE II. Evaluation Results of BM25 on the BioASQ Task B 2024 Dataset.

Metrics @10	k_1=0.9, b=0.4	k_1=0.9, b=0.8	k_1=1.2, b=0.4	k_1=1.2, b=0.8
MAP	0.6100	0.6255	0.6073	0.6263
MRR	0.8965	0.8866	0.8978	0.8886
Precision	0.5885	0.6044	0.5871	0.6088
Recall	0.4703	0.4860	0.4696	0.4884
nDCG	0.7761	0.7847	0.7744	0.7877

The results reveal that the configuration $k_1 = 1.2$, $b = 0.8$ consistently yields the best performance for most metrics. Interestingly, the highest MRR@10 is reached with $k_1 = 1.2$, $b = 0.4$, suggesting that while this setting helps retrieve the first relevant document at a higher rank, the overall ordering may be less optimal than with $b = 0.8$.

These findings emphasize the importance of hyperparameter tuning in lexical retrieval and confirm that BM25, even as a sparse model, can provide a strong and competitive baseline when properly optimized.

B. BioBERT Results

We evaluate the performance of BioBERT in two experimental settings: in its original pre-trained form and after being fine-tuned on the BioASQ TASK B 2024 training set. The results shown in Table III indicate a substantial improvement across all evaluation metrics after fine-tuning.

TABLE III. Evaluation Results of BioBERT Before and After Fine-Tuning.

Metrics	Pre-trained model	Fine-tuned model
MAP@10	0.6430	0.8180
MRR@10	0.7512	0.8883
Precision@10	0.5097	0.6088
Recall@10	0.3589	0.4884
nDCG@10	0.7578	0.8917

To further illustrate these differences, Fig. 1 presents a bar chart comparing the performance of the two models. The fine-tuned BioBERT clearly outperforms the pre-trained version, achieving a MAP@10 of 0.8180 compared to 0.6430, and an nDCG@10 of 0.8917 versus 0.7578. These results emphasize the effectiveness of domain-specific fine-tuning in enhancing ranking quality for biomedical information retrieval tasks.

To better assess the impact of dense retrieval methods in biomedical information retrieval, we compare the top-performing BM25 configuration with the fine-tuned BioBERT model. The results, summarized in Table IV, highlight the advantages of using a semantic approach over a traditional sparse retrieval model.

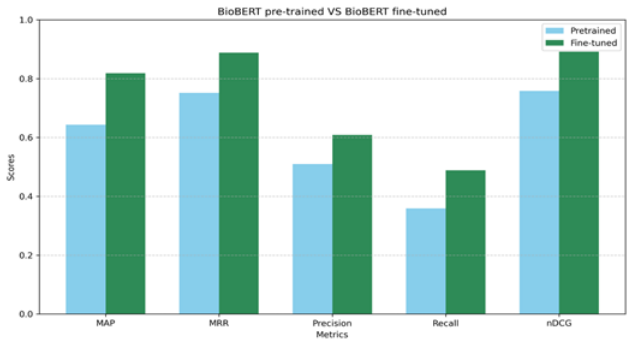

Fig. 1. Comparison of BioBERT performance in pre-trained and fine-tuned settings for each evaluation metric at rank 10.

TABLE IV. RETRIEVAL PERFORMANCE COMPARISON BETWEEN BM25 AND FINE-TUNED BIOBERT ON THE BIOASQ TASK B 2024 DATASET.

Metrics	BM25 (k_1 = 1.2, b = 0.8)	Fine-tuned BioBERT
MAP@10	0.6263	0.8180
MRR@10	0.8886	0.8883
Precision@10	0.6088	0.6088
Recall@10	0.4884	0.4884
nDCG@10	0.7877	0.8917

As illustrated in Fig. 2, BioBERT outperforms BM25 across most metrics, notably achieving improvements of +19.2% in MAP@10 and +10.4% in nDCG@10. These gains confirm the added value of contextualized language representations in capturing nuanced biomedical terminology and semantics, which are often missed by term-matching approaches like BM25.

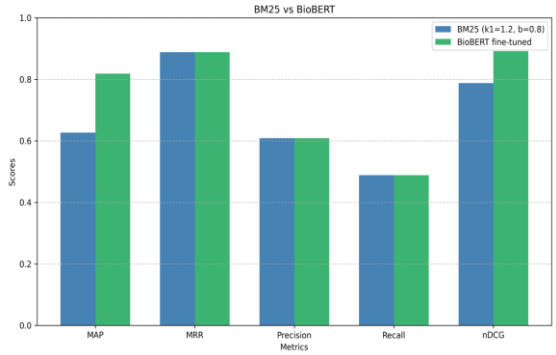

Fig. 2. Comparative performance of BM25 and fine-tuned BioBERT.

The introduction of BioBERT significantly improves ranking quality, particularly in terms of MAP@10 and nDCG@10. These gains can be attributed to the model's ability to encode contextual semantics, enabling it to capture synonymy, abbreviations, and implicit relationships that are not accessible to term-based approaches. Interestingly, while BioBERT and BM25 achieve comparable Precision@10 and MRR@10 scores, BioBERT consistently produces a more coherent ordering of relevant documents. This suggests that dense representations primarily enhance ranking quality rather than retrieval quantity, a pattern frequently observed in domain specific retrieval tasks.

C. RankFormer-Hybrid Model Results

Table V presents the performance of the proposed RankFormer-Hybrid model, summarizing results after 5 and 6 training epochs. These outcomes reflect the integration of hybrid relevance signals into the model's learning process.

TABLE V. RANKFORMER-HYBRID PERFORMANCE AT RANK 10 AFTER 5 AND 6 TRAINING EPOCHS.

Metrics / Epochs	5	6
MAP@10	0.9614	0.9526
MRR@10	0.9196	0.9118
Precision@10	0.6277	0.6242
Recall@10	0.8124	0.8081
nDCG@10	0.9320	0.9320

The model demonstrates outstanding performance across all evaluation metrics, even with a limited number of training epochs. At epoch 5, the model reaches a MAP@10 of 0.9614, MRR@10 of 0.9196, and nDCG@10 of 0.9320. These values remain stable or slightly decrease at epoch 6 due to the application of early stopping, which was employed to prevent overfitting and ensure optimal generalization.

Such high scores with minimal training suggest that the hybrid supervision significantly accelerates the convergence of the model by providing rich and informative ranking signals. This strong external guidance reduces reliance on long training schedules and enhances learning efficiency.

The rapid convergence observed within a small number of training epochs indicates that the hybrid supervision provides a stable and informative optimization signal. This behavior is particularly relevant for applied systems, where computational efficiency and training stability are critical design constraints.

These findings confirm the effectiveness of the proposed RankFormer-Hybrid model and highlight the value of incorporating complementary lexical and semantic signals in ranking models. This hybrid strategy leads to more nuanced and accurate rankings, particularly beneficial in complex systems such as biomedical information retrieval.

D. Comparative Analysis

The final comparison brings together the best-performing configurations of all evaluated models in order to highlight the progressive improvement achieved through increasingly sophisticated retrieval strategies. The results presented in Fig. 3 show a clear and consistent trend. The performance improves markedly as we move from traditional sparse retrieval to dense, and finally to learning-to-rank models.

BM25, a strong lexical baseline, performs well in precision-oriented metrics such as MAP and MRR, especially when its hyperparameters are finely tuned. However, it remains limited in its ability to capture semantic nuances, as reflected in its lower scores for recall and nDCG. The introduction of BioBERT, leads to a significant boost across all metrics. This confirms the added value of semantic understanding, particularly crucial in specialized domains such as biomedicine, where simple lexical overlap between queries and documents is often insufficient.

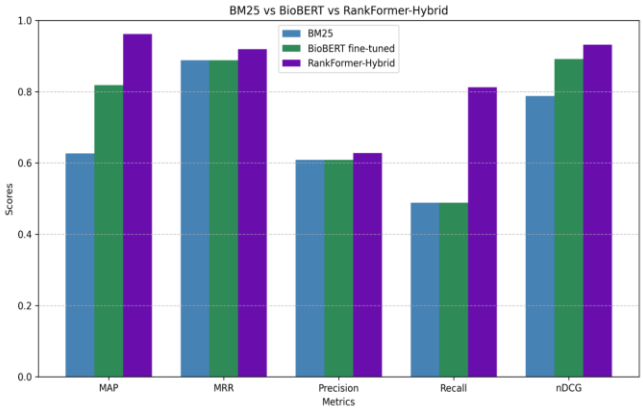

Fig. 3. Comparative performance of BM25, BioBERT, and the proposed RankFormer-Hybrid model.

The most substantial improvement is achieved by the RankFormer-Hybrid model, which combines lexical and semantic signals while leveraging supervised learning to refine document ranking. This model outperforms all others by a large margin, demonstrating the effectiveness of LTR approaches that not only understand content more deeply but also learn to prioritize relevance signals in a task-specific manner.

This comparative analysis underscores a key insight: while traditional and dense retrieval methods each contribute to performance, it is the combination of these methods within a supervised LTR framework that achieves the most robust and reliable results. In information retrieval tasks, especially those involving complex and high-stakes domains, incorporating LTR is not just beneficial but essential for achieving state-of-the-art performance.

To position our work within the current literature, we compare the performance of the proposed RankFormer-Hybrid model with that reported in selected previous studies. This qualitative comparison relies on standard evaluation metrics, including MAP, MRR, and nDCG.

Table VI presents a comparison between the proposed RankFormer-Hybrid model and other LTR models.

TABLE VI. COMPARISON OF THE PERFORMANCE OF THE PROPOSED RANKFORMER-HYBRID MODEL WITH OTHER LTR MODELS.

Model	Dataset	MAP@10	MRR@10	nDCG @10
BERT-based with injecting BM25 [27]	MS MARCO	0.367	0.364	0.422
HYRR [28]	MS MARCO	-	0.440	0.508
MultiLTR [29]	MQ2007	0.465	-	0.490
Vespa Hybrid [30]	TREC Covid	-	-	0.750
RankingRAE [31]	BioASQ Task B 2015	0.141	-	-
RankfFormer-Hybrid	BioASQ Task B 2024	0.961	0.920	0.932

The results in Table V show that the proposed RankFormer-Hybrid model outperforms several related models in terms of MAP, MRR, and nDCG at rank 10. Unlike methods that use BM25 scores as additional textual inputs or combine them through ensemble techniques, our approach directly integrates the hybrid BM25-BioBERT score into the document representation at the input level, allowing it to guide the attention mechanism from the earliest layers of the model. This integration leads to faster convergence, with the model reaching its best performance in only six epochs, which is particularly suitable for resource-constrained settings.

V. CONCLUSION

This work compares several information retrieval approaches on the BioASQ Task B 2024 dataset, ranging from the traditional BM25 (configured with k1 = 1.2, b = 0.8) to the semantic BioBERT model, and the hybrid learning-to-rank RankFormer approach. The results show a clear performance improvement as we move from sparse to dense and then to supervised hybrid methods, with RankFormer achieving the highest scores after just a few training epochs. These findings highlight the limitations of lexical-only retrieval in biomedical contexts and emphasize the effectiveness of incorporating semantic representations and hybrid signals. This study confirms the central role of learning-to-rank in modern IR systems and opens avenues for future research toward more adaptive and scalable ranking models.

A. Limitations and Discussions

While the results obtained in this study are encouraging and highlight the effectiveness of learning-to-rank techniques in biomedical information retrieval, some limitations should be acknowledged. First, the experiments are conducted exclusively on the BioASQ Task B 2024 dataset, which, although well established, may not fully represent the diversity and complexity of other biomedical or general domain IR tasks. As a result, the observed performance gains may not directly generalize to other biomedical retrieval tasks or to open-domain information retrieval settings without additional validation.

Second, although our methodology includes a hybrid scoring strategy and a deep LTR model, we limited our exploration to a single architecture and a relatively small number of training epochs. More comprehensive hyperparameter optimization and comparisons with alternative transformer-based ranking models could further strengthen the conclusions.

These limitations do not undermine the validity of the reported results; rather, they delineate the experimental boundaries of the present study and highlight directions for future research aimed at improving generalizability, robustness, and scalability.

B. Future Work

In future work, we plan to extend our evaluation to include other biomedical datasets to further assess the generalizability of our approach. Additionally, integrating user interaction signals could enable dynamic learning and continuous model improvement. Another promising direction involves exploring multi-model retrieval systems by incorporating structured knowledge or non-textual medical data into the ranking process. Finally, combining LTR with reinforcement learning or meta-learning techniques may offer new opportunities for optimizing

retrieval in highly specialized domains such as biomedical literature search.

REFERENCES

[1] H. Kulkarni, N. Goharian, O. Frieder, and S. MacAvaney, "LexBoost: improving lexical document retrieval with nearest neighbors," in ACM Symposium on Document Engineering, San Jose, CA, USA. ACM, New York, NY, USA, 10 pages, 2024.

[2] A. Goker and J. Davies, Information retrieval: searching in the 21st Century, ISBN: 9780470027622, John Wiley & Sons, Ltd, 2009.

[3] S. Sivarajkumar, H.A. Mohammad, D. Oniani, K. Roberts, W. Hersh, H. Liu, D. He, S. Visweswaran, and Y. Wang, "Clinical information retrieval: a literature review," Journal of Healthcare Informatics Research, 8(2):313–352, 2024.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Computation and Language, 2013.

[5] P.D. Turney and P. Pantel, "From frequency to meaning: vector space models of semantics," Journal of Artificial Intelligence Research, 37, 141-188, 2010.

[6] S.E. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in Information Retrieval, 3(4), 333-389, 2009.

[7] K.M. Svore and C.J.C. Burges, "A machine learning approach for improved BM25 retrieval," in Proceedings of the 18th ACM Conference on Information and Knowledge Management, HongKong, China, 2009.

[8] S.E. Robertson, H. Zaragoza, and M.J. Taylor, "Simple BM25 extension to multiple weighted fields," in Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, Washington, D.C., USA, 42-49, 2004.

[9] M. Luo, A. Mitra, T. Gokhale, and C. Baral, "Improving biomedical information retrieval with neural retrievers," Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[10] P.S. Huang, X. He, J. Gao, et al., "Learning deep structured semantic models for web search using click through data," in Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, California, USA, 2333–2338, 2013.

[11] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 373–374, 2014.

[12] J. Devlin, M.W. Chang, K. Lee,and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186, 2019.

[13] T. Brown, B. Mann, N. Ryder, and D. Amodei, "Language models are few-shot learners," 34th Conference on Neural Information Processing Systems Vancouver,Canada, 2020.

[14] A.Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

[15] J. Lee, W. Yoon, Sungdong Kim, D. Kim , Sunkyu Kim, C.Ho So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, Volume 36, Issue 4, , Pages 1234–1240, 2020.

[16] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare, 1(1), 24 pages, 2021.

[17] S. Bruch, S. Gai, and A. Ingber, "An analysis of fusion functions for hybrid retrieval," ACM Trans. Inf. Syst. 42, 1, 1-35, 2023.

[18] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, "Learning to rank using gradient descent," in Proceedings of the 22nd international conference on Machine learning, New York, NY, USA, 89–96, 2005.

[19] Z. Cao, T. Qin, T.Y. Liu, M. F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in Proceedings of the 24th international conference on Machine learning, 129–136, 2007.

[20] M. Buyl, P. Missault, and P. A. Sondag, "RankFormer: listwise learning-to-rank using listwide labels," in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3762–3773, 2023.

[21] J. Lin, X. Ma, and A. Yates, "Pretrained transformers for text ranking: BERT and beyond," Synthesis Lectures on Human Language Technologies, 14(4), 1–325, 2021.

[22] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. Alvers, D. Weissenborn, A. Krithara, S.Petridis, S.A. Pavlopoulos, P.G. Bagos, J. Bjorne, S. Pyysalo, F. Ginter, T. Salakoski, M. Schroeder, E. Gaussier, M.R. Amini, Y. Kouropalatis, C. Matuschek, L. Goeuriot, E. Gaussier, M. Stevenson, S. Ananiadou, A.Kastrin, and M. Dehmer. "An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition," BMC Bioinformatics, 16 :138, 2015.

[23] J. Lin, X. Ma, S.C. Lin, J. H. Yang, R. Pradeep, and R. Nogueira, "Pyserini: an easy-to-use Python toolkit to support replicable IR research with sparse and dense representations," in proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.

[24] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E. M. Voorhees, "Overview of the TREC 2019 deep learning track," in TREC 2019 Conference Notebook. National Institute of Standards and Technology, 2020.

[25] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, ISBN: 0262035618, 2016.

[26] S. Chavhan, M. Raghuwanshi, and R.C. Dharmik, "Information retrieval using machine learning for ranking: a review," Journal of Physics Conference Series 1913(1):012150, 2021.

[27] A. Askari, A. Amin, P. Gabrielle, K. Wessel, and V. Suzan, "Injecting the BM25 score as text improves BERT-based re-rankers," in proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), pages 1682–1686, 2023.

[28] J. Lu, H. Keith, B. Oguz, M. Ji, and N. Jianmo, "HYRR: Hybrid infused reranking for passage retrieval," in proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 139–151, 2022.

[29] H. Yang and T. Gonçalves, ""MultiLTR: Text ranking with a multi-stage learning-to-rank approach," Information, 16(4) :308, 2025.

[30] J. K. Bergum, "Improving zero-shot ranking with vespa hybrid search – part two," https://blog.vespa.ai/improving-zero-shot-ranking-with-vespa-part-two/, 2023. Vespa Blog.

[31] Y. Yan, B. Zhang, and L. Xu-Feng, "RankingRAE : Global learning to rank for question answering with recursive autoencoders," PLOS ONE, 15(11) :e0242061, 2020.