Generative Large Language Models Augmented Hybrid Retrieval System for Biomedical Question Answering⋆

Notebook for the BioASQ Lab at CLEF 2024

Bor-Woei Huang[1]

[1]University of Padova, Italy

Abstract

In our participation in the 12th BioASQ challenge document retrieval task B Phase A, our system hybridized advanced generative large language models (LLMs), traditional BM25 lexical retriever, and BERT-based cross-encoder re-ranker. Our approach aimed to improve recall and MAP scores by utilizing the pseudo-documents generated by GPT-3.5 and Gemini to augment a hybrid two-stage retrieval system. Our retrieval system is composed of two modules: the generator module and the extractor module. The generator module consists of LLMs GPT-3.5 and Gemini, which generate synthetic data on the seen questions and pseudo-documents on the unseen test questions. The extractor module has a two-stage retrieval framework consisting of BM25 lexical retriever and BiomedBERT cross-encoder re-ranker. By expanding the original queries with the biomedical entities extracted from the pseudo-documents, we observed improvements in recall in the first stage BM25 retrieval. We explored the effectiveness of a BiomedBERT cross-encoder re-ranker trained on the data that combines golden-standard data, synthetic data and LLM-generated pseudo-documents for the second stage re-ranking. This approach improves MAP scores by considering the contextual relationships between test questions and pseudo-documents. Finally, we used fusion methods to optimize the final output considering the question types. Our system demonstrated promising results in effectively retrieving relevant biomedical documents from PubMed.

Keywords

Zero-shot, Generative large language models, BioASQ, Question Answering

1. Introduction

We present our participating system in the document retrieval task B phase A at the 12th BioASQ challenge. We applied LLMs to augment the performance of a hybrid two-stage retrieval system. First, for each unseen test question, we utilized GPT-3.5 and Gemini to generate pseudo-documents containing information relevant to the question. Following this, we implemented biomedical named entity recognition tools (BioNER) to extract biomedical entities from these pseudo-documents. The biomedical entities were used to improve BM25 retrieval, while the pseudo-documents were used to improve the BERT cross-encoder re-ranker. In the first retrieval stage, we expanded the queries with the biomedical entities by concatenating them with the original queries for the BM25 searcher, enhancing the recall capability of this lexical retrieval method. In the second stage, we employed a BiomedBERT cross-encoder re-ranker. This re-ranker was trained on a combination of golden standard data, synthetic data, and LLM-generated pseudo-documents that correspond to the test questions that provide context and domain-specific information. Last but not least, depending on the question types, we applied the Reciprocal Rank Fusion (RRF) method to fuse the outputs from different stages of the systems with and without the assistance of LLMs. It helps to leverage the strengths of different retrieval methods and enhance overall performance. An overall architecture of our framework is presented in Figure 1.

2. Related Works

The primary goal of the challenge of BioASQ-12B Phase A [1] is to assess the ability of various systems to retrieve and rank the most relevant biomedical articles in response to a given set of questions. Each

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
borwoei.huang@studenti.unipd.it (B. Huang)

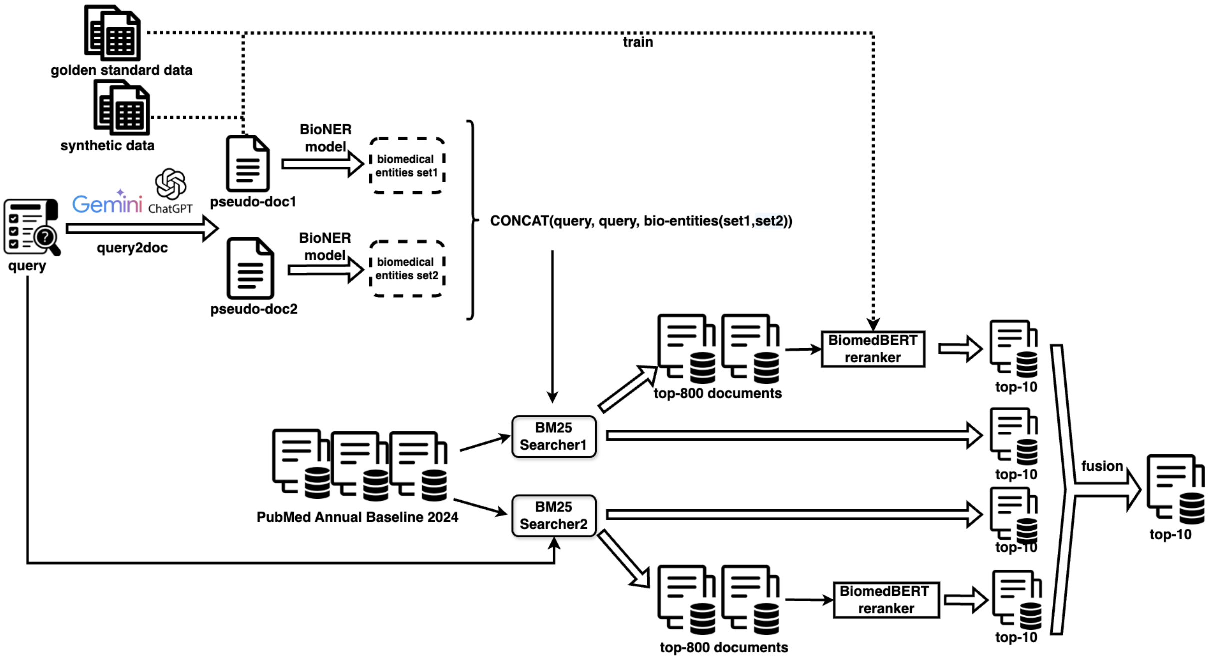

Figure 1: Overview of the document retrieval system pipeline.

test batch in Phase A consists of 85 biomedical questions. The task requires, for each question, ranking up to 10 of the most relevant biomedical articles from the PubMed Annual Baseline 2024 database[1], which is a database snapshot of the PubMed bibliographic repository[2]. This database includes all PubMed citations from the year 2024, as well as any updates made to older citations. The Mean Average Precision (MAP) metric is used to evaluate the performance of the participating systems.

There are four types of questions given in the challenge: factoid, list, summary, and yes/no. The factoid-type questions ask for certain factual information, like the diagnosis of a disease or the role of a protein/gene. The list-type questions look for a list of objects related to a particular topic, such as the symptoms of a disease. The summary-type questions look for a comprehensive summary of a specific topic or concept, such as the mechanism of action of a drug or the pathophysiology of a disease. The yes/no type questions require a true or false response to a statement.

2.1. Generator module

Large Language Models for zero-shot problem: GPT-3 is a lightweight, open-source generative language model. Unlike larger models such as GPT-3.5 or Llama 3, GPT-3 is designed to run efficiently on standard personal laptops. Trained on a vast corpus of biomedical literature, including PubMed articles and clinical trial data, BioGPT is adept at generating coherent and contextually relevant text within the biomedical field. However, it's worth noting that the latest version of BioGPT was last updated in early 2023, so it may lack the most current biomedical knowledge from 2023 and 2024, potentially leading to deviations from accurate answers to questions pertaining to those years.

In contrast, ChatGPT, based on the GPT-3.5 architecture, is widely recognized as one of the leading generative AI models. While it's true that "gpt-3.5-turbo" isn't the most recent iteration of GPT[3], it still holds significant value due to its extensive pre-training and capabilities. Despite having a knowledge cutoff set at September 2021, GPT-3.5 offers impressive performance and versatility across various tasks. Additionally, it's worth noting that GPT-3.5 tends to be more cost-effective compared to newer models, making it a practical choice for many applications where the latest updates may not be essential.

[1] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://platform.openai.com/docs/models

On the other hand, Gemini Ultra 1.0 offers a compelling advantage with its knowledge cutoff extending to early 2024[4], which can be particularly beneficial for accessing up-to-date information on biomedical topics, especially for biomedical questions from 2023 and 2024. Additionally, the fact that Gemini Ultra 1.0 is available free of charge enhances its appeal, as it provides access to state-of-the-art biomedical information retrieval capabilities without any financial burden. This combination of current knowledge and cost-effectiveness makes Gemini Ultra 1.0 an attractive option for researchers and anyone seeking accurate biomedical information.

2.2. Extractor module

Hybrid two-stage retrieval [2] is a popular approach in information retrieval (IR) that takes advantage of the strengths of different retrieval methods to improve overall search performance. It typically involves a sparse retrieval step to generate a larger set of potential documents, followed by a re-ranking step to select the most relevant documents from the candidates. Lexical retrievers are commonly used in the first retrieval stage for their effectiveness. Okapi BM25 [3] is particularly often implemented, relying on term frequency and inverse document frequencies to match queries to documents. Transformer-based re-ranker is popularly used in the second stage retrieval for their high precision that directly compares the similarity of a pair of questions and potential answers. This lexical retrieval and BERT re-ranking two-stage framework has been widely adopted in various applications [4].

First-stage BM25 retrieval BM25 is a ranking algorithm in IR that estimates the relevance of documents to a given search query. BM25 relevance score is defined as:

$$\text{Relevance Score (D, Q)} = \sum_{q_i \in Q} \text{IDF}(q_i) \cdot \frac{tf(q_i, D) \cdot (k_1 + 1)}{tf(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \tag{1}$$

where IDF(q_i) is the inverse document frequency weight of the query term q_i, and $tf(q_i, D)$ is the frequency of a query term q_i that appears in the document D. $|D|$ is the length of the document, D and avgdl is the average document length in the collection. The higher the term frequency within the document D, the more relevant it is. On the other hand, IDF(q_i) suppresses the influence of uninformative terms such as "the", "of" and "and". The rare terms relevant to the query are given more weight.

Second-stage re-ranking Using a BERT-based cross-encoder [5] to capture the contextual relationship between a question and an answer is a common approach for re-ranking the candidate documents to improve precision. Cross-encoder tends to yield more accurate relevance scores than bi-encoder models, as it considers the full context of both the question and answer. Unlike bi-encoder models, which independently encode the question and answer and then compute their similarity, cross-encoder models jointly encode both the question and answer together, allowing for a more comprehensive understanding of their relationship. By concatenating the question and answer candidate as input to the cross-encoder, the model learns to assess the answer's relevance in the context of the given question. The sequences of query terms and document words are joined with the [SEP] token, and the BERT cross-encoder computes a relevant score for the representation of the [CLS] token with linear layer W_s:

$$\text{Score} = \text{cross-encoder([CLS] query [SEP] abstract [SEP])} * W_s \tag{2}$$

The relevance score indicates the likelihood that the answer is pertinent to the question.

2.3. LLM augmented two-stage IR systems on zero-shot tasks

Query expansion with LLMs The traditional approach to query expansion involves identifying synonyms and related terms in the query to retrieve more relevant documents and improve search result

[4]https://ai.google.dev/gemini-api/docs/models/gemini

recall [6]. Document expansion extends this concept by expanding the query terms within documents [7]. However, when queries are brief or imprecise, the retrieved documents may not align well with the original query. State-of-the-art generative AI models now offer reliable inherent knowledge for zero-shot information retrieval systems. Query expansions generated by LLMs surpass traditional methods by using the expansive knowledge embedded within these language models, such as ChatGPT or Gemini, without requiring fine-tuning [8, 9]. While similar to previous works [9, 10], our approach is distinctive in focusing on prompting LLMs, and then expanding the query with the extracted biomedical entities rather than entire LLM-generated pseudo-documents because the key terms determining document relevance are the biomedical entities found within documents in many biomedical information retrieval tasks. Therefore, our approach expands queries with biomedical named entities extracted from the LLM-generated pseudo-documents.

Re-ranker trained on LLM-generated data Askari et al. [11, 12] have studied the effectiveness of BERT cross-encoder re-ranker fine-tuned on the data generated with GPT-3.5 (ChatGPT). The results indicate that a cross-encoder trained on the LLMs generated data is reliable for retrieving unseen human written documents on the seen queries. Inspired by this work, we trained our BiomedBERT cross-encoder re-ranker on the GPT-3.5 and Gemini-generated text (pseudo-documents) on the test questions to essentially convert the "unseen" test queries to "seen" queries from the perspective of the cross-encoder re-ranker. The LLM-generated pseudo-documents can broaden the scope of the re-ranker's knowledge, potentially leading to retrieving relevant documents that might have been missed. However, the effectiveness of this method hinges on the quality and relevance of the LLM-generated pseudo-documents, as LLMs can sometimes generate biased, inaccurate, or redundant information.

3. Methodology

3.1. Pseudo-documents and synthetic data

We prompt GPT-3.5 turbo and Gemini Ultra 1.0 to generate synthetic data and pseudo-documents for the biomedical questions. The prompt formulated is simply

LLM prompts: Write a paragraph that answers {query}

LLMs were prompted to create synthetic data based on previous years' BioASQ questions. Negatively labeled data are randomly sampled from documents not positively labeled for a given query. This results in a large set of (query, synthetic answer) pairs, which we combine with the gold standard dataset [13] to form our training data. GPT-3.5 excels at generating well-written, concise text, but its knowledge cutoff might be a bit older, so it may not always have the most up-to-date information on rapidly evolving biomedical topics. In contrast, Gemini can access the latest information, which is essential for staying current in the dynamic biomedical field. The documents generated by ChatGPT are typically longer than those from Gemini, as ChatGPT tends to provide extra explanations on proper nouns and terminologies that might not directly answer the queries. On the other hand, Gemini's shorter answers usually more directly address the questions. While the generated documents are generally accurate, they can occasionally produce factually inaccurate content. We may obtain more accurate pseudo-documents by balancing the detailed explanations from ChatGPT with the concise, direct responses from Gemini. This hybrid approach enhances the quality and relevance of the synthetic data and pseudo-documents.

3.2. First-stage retriever: BM25

In the first-stage of document retrieval, we implemented Pyserini [14], a Python wrapper for Anserini's Java-based Apache Lucene [15], that provides tools to index large collections of documents using the

BM25 IR model and search over the indexes. We retrieved the top 800 highest BM25 scoring documents from the PubMed Annual Baseline 2024. The query expansion by prompting LLMs technique [10] is an effective way of improving the performance of BM25 by incorporating the pseudo-document generated by ChatGPT and Gemini. Unlike Wang et al. [10], we didn't use the full pseudo-documents for BM25 retrieval because they contain many non-crucial terms, which can lead to many false positives. We then implemented the SciSpacy named entity recognition tool to identify the biomedical and clinical name entities in them. Table 1 is an example that demonstrates the extracted biomedical entities from ChatGPT and Gemini-generated documents and their high lexical overlapping with the ground truth answers. BM25 strongly relies on the query terms in the questions. The more specific the words in the question, the higher the recall the BM25 search retriever can achieve. The identified biomedical and clinical entities in the pseudo-documents are the critical words for linking the questions with the targeted documents. Therefore, by appending the biomedical entities with the original questions, we can recall more correct documents in the first stage retrieval.

Furthermore, to ensure that the terms in the original questions have sufficient influence on the BM25 relevance score. We increased the weight of the original query terms by repeating them multiple times, followed by concatenating the biomedical entities extracted from the pseudo-documents. The new query is

$$\text{new query } q' = Concat(q, q, BioEntities(set1, set2)) \tag{3}$$

where q is the original query, which is repeated to increase their relative weights with respect to the extracted biomedical entities. "Concat" is the string concatenation operator. "BioEntities" set1 and set2 are the biomedical and clinical entities extracted from the GPT-3.5 and Gemini-generated pseudo-documents, respectively.

Specifically, we repeated the original queries four times for the batch 2 test and two times for the batch 3 test. In contrast, for the batch 4 test, we did not repeat the original query terms to magnify the influence of the generative LLMs. This approach allowed us to evaluate the impact of emphasizing original query terms versus the additional biomedical entities provided by the LLMs on the performance of the BM25 retrieval method. By adjusting the repetition of the original queries, we can understand and balance the influence of the original terms and the enriched context provided by the LLM-generated pseudo-documents.

3.3. Second-stage: cross-encoder re-ranker

We fine-tuned a BERT-based cross-encoder re-ranker in two stages to improve its performance. In the first stage, we fine-tuned a pre-trained BiomedBERT model using a combination of gold standard data and synthetic data. The use of synthetic data is to augment gold standard data and provide the cross-encoder with a wider range of examples. During this stage, the model learned to predict the relevance score for each question-answer pair. This process allowed the cross-encoder to capture intricate relationships between questions and answers. In the second stage, we conceptually transferred the zero-shot task to a pseudo-few-shot task by training the BiomedBERT re-ranker on the LLM-generated pseudo-documents using the test queries on the test days. Few-shot learning tends to outperform zero-shot learning in various tasks. Zero-shot learning heavily depends on knowledge gained during a golden standard data training phase to infer relationships between seen and unseen queries. Zero-shot learning can struggle if the unseen question is conceptually distant from what the model was trained on. In our pseudo-few-shot learning, the model gets pseudo-documents from ChatGPT and Gemini AI models for each test question. The knowledge inherent in the LLM-generated text helps the BERT re-ranker quickly adapt to new question tasks. These pseudo-documents, crafted with information pertinent to the test queries, provided the re-ranker with specific contextual examples.

Table 1
An example of ChatGPT and Gemini-generated documents and the extracted biomedical entities on a biomedical question. Comparing these to the ground truth answers. Bold texts are the overlapping terms between the extracted biomedical and clinical entities and ground truth answers.

query	What type of drug is administered to the mother for the prevention of lung immaturity when a preterm delivery is anticipated?
ChatGPT generation	When a preterm delivery is anticipated, corticosteroids such as betamethasone or dexamethasone are administered to the mother for the prevention of lung immaturity in the premature infant. These corticosteroids stimulate the production of surfactant in the fetal lungs, which plays a crucial role in reducing surface tension and preventing alveolar collapse during respiration. By enhancing surfactant production, corticosteroid administration helps improve lung compliance and function in preterm infants, reducing the risk of respiratory distress syndrome (RDS) and other complications associated with prematurity. The timing and dosing of corticosteroids are carefully determined based on gestational age and anticipated timing of delivery to maximize the benefits of lung maturation while minimizing potential adverse effects.
Gemini generation	When preterm delivery is anticipated, corticosteroids are typically administered to the mother to accelerate fetal lung maturation and reduce the risk of respiratory distress syndrome (RDS) in the newborn. These drugs, such as betamethasone or dexamethasone, mimic the natural cortisol surge that occurs before birth, stimulating the production of surfactant, a substance crucial for proper lung function. This intervention has proven to be highly effective in reducing neonatal morbidity and mortality associated with preterm birth.
extracted biomedical entities	ChatGPT{betamethasone, dexamethasone, lung immaturity, corticosteroids, surfactant, respiratory distress syndrome, RDS, prematurity, lung, fetal lungs, surface, alveolar, corticosteroid} , Gemini{corticosteroids, respiratory distress syndrome, RDS, betamethasone, dexamethasone, cortisol, fetal lung, lung }
ground truth	"The administration of antenatal **corticosteroids**, in cases of imminent preterm delivery, can enhance **fetal lung** maturation and reduce the incidence of **respiratory distress syndrome**, leading to improved neonatal outcomes.", "single course of antenatal corticosteroids from 24 up to 34 gestational week", "use of antenatal **corticosteroids** to prevent **respiratory distress syndrome** (**RDS**) in preterm infants", "beneficial effect of antenatal **corticosteroids** on **fetal lung** maturation and widespread recommendations to use this treatment in women at risk of preterm delivery" , "course of **corticosteroids** to women prior to anticipated preterm birth (before 37 weeks of pregnancy)", "use of a single course of antenatal **corticosteroids** to accelerate **fetal lung** maturation in women at risk of preterm birth"

3.4. Fusion

The LLM-augmented retrieval method heavily relies on the quality of the LLM-generated pseudo-documents. The LLM-generated text is potentially inaccurate or overly informative, which can lead to decreased performance on some types of questions. While traditional BM25 retrieval and LLM-augmented BM25 retrieval rely on lexical matching, the BiomedBERT cross-encoder captures semantic relationships and contextual meaning. To balance the strengths and weaknesses, fusing the results of different stages of the system with and without the assistance of LLMs is a mitigation strategy to optimize the final output wisely. Our submitted runs are the reciprocal rank fusion (RRF) [16] of four ranks with a higher weight on the more recent years PubMed abstracts. The four ranks we took from our system are:

- **BM25LLM**: the output of the LLM-augmented BM25 searcher
- **BM25LLM/re-ranker**: the output from the retrieval of the LLM-augmented BM25 searcher followed by the cross-encoder re-ranker (LLM-augmented two-stage retrieval)
- **BM25**: the output of standard BM25 searcher

- **BM25/Re-ranker**: the output from the retrieval of the standard BM25 searcher followed by the cross-encoder re-ranker (standard two-stage retrieval)

The RRF scores are computed as follows:

$$RRF(d) = \sum_{r \in R} \frac{1}{k + r(d)} \tag{4}$$

where $r(d)$ is the position of the document, and we set the constant k to 0.

Note that the output of RRF is not always better than either one of the results of the four methods. It can be worse than the best result of the four ranks if including a low-performing ranked list in the fusion process, which dilutes the effectiveness of the best-performing one. For this reason, we would like to choose to fuse the two results with the highest recall for each question type so that we optimally take advantage of the BM25 retrievers and cross-encoder re-ranker. We have experimented with several fusion combinations. The question-type dependent fusion combinations that we have experimented with are listed in Table 2.

Table 2
A list of fusion combinations we have tried.

fusion	Question types			
	list	factoid	summary	yes/no
1	BM25/re-ranker	BM25/re-ranker	BM25/re-ranker	BM25/re-ranker
2	BM25LLM/re-ranker	BM25LLM/re-ranker	BM25LLM/re-ranker	BM25LLM/re-ranker
3	BM25/re-ranker	BM25LLM/re-ranker	BM25LLM/re-ranker	BM25/re-ranker
4	BM25LLM/re-ranker	BM25/re-ranker	BM25LLM/re-ranker	BM25LLM/re-ranker
5	RRF(BM25,re-ranker)	BM25LLM/re-ranker	RRF(BM25LLM,re-ranker)	BM25/re-ranker
6	RRF(BM25LLM,re-ranker)	BM25LLM/re-ranker	BM25LLM/re-ranker	BM25/re-ranker
7	RRF(BM25,re-ranker)	RRF(BM25LLM,re-ranker)	BM25LLM/re-ranker	BM25/re-ranker
8	RRF(all 4 ranks)	RRF(all 4 ranks)	RRF(all 4 ranks)	RRF(all 4 ranks)

4. Results and Discussion

For the first retrieval stage, the effectiveness of the biomedical entities on the term matching can be seen in Table 3. The benefits and disadvantages of the LLM-generated biomedical terms on the BM25 retriever vary depending on the types and structures of the questions. The LLM-generated pseudo-document documents systematically decreased the recall of the summary-type questions while increasing the recall of the factoid-type questions across three batches.

For test batch 2, the original query terms were repeated four times, so the influence of the LLM-generated biomedical entities was not prominent. For test batch 3, the original query terms were repeated twice, and the overall recall was highest, suggesting a good balance between original query terms and LLM-generated terms. For test batch 4, we appended the extracted biomedical entities to the original queries. The strongest impact of LLM-generated biomedical terms was seen in this batch, where they were given a higher weight. The impact of the biomedical terms is stronger in test batch 4. Although the overall recall decreased, we observed improvement in the factoid and yes/no-type questions.

The disadvantage of the BM25 ranking algorithm is that it is incapable of capturing the semantics and parsing the syntax of the questions and documents, so the BM25 searcher doesn't understand the meaning of the questions and answers. BM25 algorithm can miss the documents that do not contain less frequent query terms but provide the correct answer to the question. For example, for the question "What is borderline personality disorder?" the PMID2886495 abstract discussing the treatment of patients with borderline personality disorder has a high BM25 score as "borderline personality disorder" occurs many times in the abstract, even though it doesn't answer the question. Instead, the PMID29795363 abstract addresses the syndromes of borderline personality disorder, which answers the question but does not frequently mention "borderline personality disorder," leading to a low BM25 similarity score. However, if the clinical term "impulsive behavior" in the pseudo-documents was captured by BioNER, it could help BM25 retrieve the PMID29795363 abstract that had been missed with the original query.

Table 3
The evaluation of stage-one BM25 retrieval. The notation used here: BM25 is the standard Okapi BM25 ranking algorithm, and BM25_LLM is the LLM-augmented BM25 method.

Batch	model	Recall@800	Recall			
			list	summary	factoid	yes/no
2	BM25	0.63	0.5	0.806	0.5	0.797
	BM25_LLM	0.63	0.5	0.758	0.56	0.797
3	BM25	0.6933	0.645	0.756	0.75	0.806
	BM25_LLM	0.769	0.64	0.71	0.827	0.836
4	BM25	0.639	0.595	0.723	0.674	0.622
	BM25_LLM	0.602	0.5	0.5957	0.721	0.66

The challenge of redundant query expansion underlines the careful balance needed in making use of large generative AI models for information retrieval tasks. While these models can generate rich pseudo-documents containing a wide range of relevant information, they may also introduce non-essential details that could potentially confuse traditional term-matching retrieval methods. In particular, the BM25 retriever may struggle with the increased noise introduced by excessive expansion terms, leading to decreased relevance scores for actual correct documents. The BERT re-ranker trained on these pseudo-documents may be susceptible to being misled by redundant information, further complicating the re-ranking process.

Moreover, there's a discrepancy between the content of LLM-generated pseudo-documents and PubMed articles. The pseudo-documents often synthesize information from multiple sources, potentially spanning different PubMed abstracts. Consequently, a single correct PubMed abstract may not necessarily receive the highest relevance score, as the required information might be dispersed across multiple sources. This issue is exacerbated in list or summary type questions where the answer to a query comprises a collection of diverse documents, where individual correct documents may not receive optimal relevance scores, leading to potential oversights in term matching retrieval.

Table 4
Evaluation of the top 10 retrieved passages in the second stage retrieval with and without the assistance of LLMs.

Batch	model	MAP@10	Recall@10			
			list	summary	factoid	yes/no
2	standard two-stage retrieval	0.163	0.2	0.274	0.2	0.406
	LLM-augmented two-stage retrieval	0.1824	0.22	0.387	0.26	0.49
3	standard two-stage retrieval	0.1268	0.39	0.24	0.134	0.253
	LLM-augmented two-stage retrieval	0.1367	0.375	0.219	0.154	0.254
4	standard two-stage retrieval	0.177	0.095	0.27	0.34	0.3
	LLM-augmented two-stage retrieval	0.182	0.095	0.191	0.44	0.32

The recall@800 in the first-stage retrieval directly influences the output of the downstream second-stage re-ranker. In the re-ranking stage, generally speaking, the types of questions with higher recall in the first stage retrieval tend to have higher recall in the second stage as well (Table 4). However, there are exceptions, such as the summary-type questions in batch 2, where this correlation does not hold as strongly. Overall, the MAP@10 is improved with the augmentation of ChatGPT and Gemini LLMs. These LLMs help enhance the performance of both the first-stage BM25 retrieval and the second-stage BERT cross-encoder re-ranking, leading to better retrieval accuracy and relevance.

4.1. Question type analysis

For the list-type questions, we initially anticipated that the LLM-assisted BM25 searcher would highly improve the recall and MAP scores. This expectation was assuming that the clinical and biomedical entities present in the LLM-generated pseudo-documents would likely well-match the answers to the list-type questions. However, the test results indicated that some of the biomedical entities in the pseudo-documents were not in the correct PubMed abstract to be retrieved, which can negatively impact the performance of the LLM-assisted BM25 retrieval in some cases, resulting in worse output compared to the standard two-stage retrieval approach. For example, for test batch 3, the recall of our LLM-assisted system decreased from 0.39 to 0.375 on the list-type questions.

For the summary-type question, some LLM-generated documents are inaccurate or cover too much extra information, negatively impacting precision. Our LLM-augmented two-stage retrieval system's recall@10 (0.219) is lower than the standard two-stage retrieval's Recall@10 (0.24) in the batch 3 test. An example illustrating this issue is the ChatGPT-generated answer to the summary question, "What was tested in the PATCH-Trauma trial?" Instead of solely identifying *tranexamic acid*, the generated answer provided extensive details about what *tranexamic acid* is, deviating from the main question and resulting in missed correct documents. This extraneous information led to poorer performance in the LLM-assisted BM25 retrieval, which subsequently decreased the second-stage re-ranking's final MAP. Moreover, the answers to the summary-type questions are typically explanatory and don't rely on the lexical overlap. The performance of our system on summary-type questions was notably affected by the pseudo-documents generated by the LLM in batch 4 testing. This issue was particularly evident for non-traditional question formats that deviated from the typical "What is ...?" or "Why does ...?" structure. Our BERT re-ranker demonstrated subpar performance when confronted with these unconventional question forms. For instance, questions like "Circular RNAs (circRNAs) are a distinct family of RNAs, how are they derived?", or indeterminate questions like "Transgender identity and mortality.", and "Burkitt lymphoma and Epstein-Barr virus." posed challenges for our re-ranker due to their deviation from standard question formats. This highlights a need for improvement in handling diverse question structures to enhance the overall robustness of our system across different question structures.

For factoid-type questions, the LLM-augmented two-stage retrieval system consistently achieved higher MAP@10 and recall@10 than the standard two-stage retrieval system. The inclusion of biomedical entities from pseudo-documents significantly boosts the BM25 relevance score of certain correct articles, and the re-ranker's precision is enhanced by training on these pseudo-documents. For example, for the question "Are children affected by multiple sclerosis?", the key biomedical entity "Pediatric MS" extracted from the pseudo-document aids in retrieving document PMID32940341, which is missed by both the standard BM25 searcher and the BERT re-ranker. Therefore, the optimal strategy for factoid-type questions is to either fuse the results of the LLM-augmented BM25 and BM25LLM/re-ranker or directly use the output of BM25LLM/re-ranker. This approach ensures that relevant documents are not overlooked, thus improving overall retrieval performance.

For yes/no-type questions, the LLM-augmented BM25 retrieval method performs only marginally better than the standard BM25 method due to the lack of decisive key biomedical entities in the answers. The primary challenge with yes/no-type questions lies in the fact that correct documents often contain statements supporting or contradicting the question without explicitly mentioning "YES" or "NO." Despite this, the descriptions of yes/no-type questions are generally less ambiguous compared to other question types, which is helpful for the second-stage BERT cross-encoder re-ranking. Effectively retrieving the correct documents for these questions requires parsing the semantic and syntactic structures of both the questions and the answers, necessitating reliance on the BERT cross-encoder re-ranker rather than solely on lexical overlap. The precise syntax and semantics in the descriptions of yes/no type questions enable the BERT re-ranker model to link the questions to their corresponding answers accurately. For instance, consider the yes/no question "Is there a specific cure for Ehlers-Danlos Syndrome?" The correct answer, "The management of this disease is possible; however, there is no cure as of present," does not contain any critical biomedical entities that directly link the question to the answer. Instead, it requires understanding the meaning behind the words to infer the response correctly.

Therefore, the output of the LLM-augmented two-stage system is optimal for yes/no type questions, minimizing the need for RRF.

4.2. The optimal fusion method

After re-examining the runs we submitted (Table 5), we realized that some fusion results turned out to be even worse than the non-fusion ones. This indicates that our approach to combining the outputs from different stages and methods was suboptimal, leading to a decrease in overall performance. In future iterations, we need to refine our fusion strategy to ensure that we are effectively leveraging the strengths of both traditional and LLM-augmented retrieval methods. Thus far, we have found that the best fusion method is:

- list: RRF(BM25LLM,re-ranker) or BM25LLM/re-ranker
- summary: RRF(BM25,re-ranker)
- factoid: RRF(BM25LLM,re-ranker) or BM25LLM/re-ranker
- yes/no: BM25LLM/re-ranker

where, for the LLM-augmented BM25 query expansion, the original query terms were repeated twice and appended to the biomedical entities extracted from the LLM-generated pseudo-documents.

Table 5
The evaluation of the best of our submissions in each test batch.

Batch	Evaluation		submission rank	best of our submissions	
	MAP	recall		question type	fusion
2	0.182	0.338	13/53	list	BM25LLM/re-ranker
				summary	BM25LLM/re-ranker
				factoid	BM25LLM/re-ranker
				yes/no	BM25LLM/re-ranker
3	0.134	0.210	24/58	list	RRF(BM25LLM,re-ranker)
				summary	BM25LLM/re-ranker
				factoid	BM25LLM/re-ranker
				yes/no	BM25/re-ranker
4	0.144	0.259	26/49	list	RRF(BM25,re-ranker)
				summary	RRF(BM25LLM,re-ranker)
				factoid	RRF(BM25LLM,re-ranker)
				yes/no	BM25/re-ranker

5. Conclusion

With the augmentation of ChatGPT and Gemini, we have demonstrated that the two-stage retrieval system shows a notable improvement in MAP@10 and recall@10. While not all types of questions benefit equally from the inclusion of LLM-generated content, the overall MAP@10 has increased. This improvement highlights the potential of LLMs to enhance retrieval performance, particularly when their outputs align well with the queries.

We recognize that LLMs' influence can be negative for certain question types. In these cases, the LLM-generated pseudo-documents may introduce noise or irrelevant information, which can decrease the retrieval process's effectiveness. To mitigate this, we have investigated fusion methods that selectively adopt the optimal results from different methods' outputs. This approach ensures that we leverage the strengths of both traditional retrieval methods and LLM-assisted approaches while minimizing the potential downsides.

For future work, experimenting with various prompt formulations may also help generate more effective responses from LLMs. Therefore, optimizing pseudo-document generation by investigating different prompt engineering techniques or exploring more advanced LLM-augmented approaches could lead to more informative and relevant pseudo-documents. Currently, our fusion method based on question type is naive and not the most effective strategy. We acknowledge the need for more sophisticated fusion methods to leverage better the strengths of both traditional and LLM-assisted retrieval systems.

We emphasize that the quality of LLM-generated pseudo-documents can vary depending on the questions' structure. This variability can sometimes reduce the recall of BM25 and the accuracy of BERT re-ranker when the pseudo-document content includes irrelevant, biased or redundant information. This issue can be eased by expanding the retrieval system from the current two modules (i.e., generator and extractor) to three modules by adding a discriminator module to supervise and improve the quality of the generated pseudo-documents, filtering out less useful content and ensuring that only the most relevant and accurate information is used in the training and retrieval process.

References

[1] A. Nentidis, A. Krithara, G. Paliouras, M. Krallinger, L. G. Sanchez, S. Lima, E. Farre, N. Loukachevitch, V. Davydova, E. Tutubalina, Bioasq at clef2024: The twelfth edition of the large-scale biomedical semantic indexing and question answering challenge, in: European Conference on Information Retrieval, Springer, 2024, pp. 490–497.

[2] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading wikipedia to answer open-domain questions, arXiv preprint arXiv:1704.00051 (2017).

[3] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University, Springer, 1994, pp. 232–241.

[4] A. Ceccato, L. Fabbian, B.-W. Huang, I. U. Khan, H. Singh, N. Ferro, et al., Seupd@ clef: Team hiball on incremental information retrieval system with rrf and bert, in: CEUR WORKSHOP PROCEEDINGS, volume 3497, CEUR-WS, 2023, pp. 2396–2415.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] C. Carpineto, G. Romano, A survey of automatic query expansion in information retrieval, Acm Computing Surveys (CSUR) 44 (2012) 1–50.

[7] M. Efron, P. Organisciak, K. Fenlon, Improving retrieval of short texts through document expansion, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 911–920.

[8] V. Claveau, Neural text generation for query expansion in information retrieval, in: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2021, pp. 202–209.

[9] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, M. Bendersky, Query expansion by prompting large language models, arXiv preprint arXiv:2305.03653 (2023).

[10] L. Wang, N. Yang, F. Wei, Query2doc: Query expansion with large language models, arXiv preprint arXiv:2303.07678 (2023).

[11] A. Askari, M. Aliannejadi, E. Kanoulas, S. Verberne, A test collection of synthetic documents for training rankers: Chatgpt vs. human experts, in: The 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023), 2023.

[12] A. Askari, M. Aliannejadi, E. Kanoulas, S. Verberne, Generating synthetic documents for cross-encoder re-rankers: A comparative study of chatgpt and human experts, in: Generative Information Retrieval workshop at ACM SIGIR 2023, 2023.

[13] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, Bioasq-qa: A manually curated corpus for biomedical question answering, Scientific Data 10 (2023) 170.

[14] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: Proceedings

of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2356–2362.

[15] P. Yang, H. Fang, J. Lin, Anserini: Enabling the use of lucene for information retrieval research, in: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, 2017, pp. 1253–1256.

[16] G. V. Cormack, C. L. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 758–759.