LLM-Guided RAG: A Biomedical Retrieval System for BioASQ Synergy 2026

BioASQ Synergy Task 14 Working Notes at CLEF 2026

Sukprasan Rattanapani[1,*], Daniela Raicu[1], Jacob Furst[1] and Alexandru Iulian Orhean[1,*]

[1]Jarvis College of Computing and Digital Media, DePaul University, Chicago, IL, USA

Abstract

Biomedical literature retrieval is challenging because scientific questions often contain specialized terminology, abbreviations, biomedical entities, and complex relationships that require evidence-grounded search. To address these challenges, we propose the design and implementation of an LLM-guided retrieval-augmented biomedical retrieval system. The proposed system draws inspiration from an ongoing unpublished project that aims to provide discovery and search services for genomic and proteomic data. More specifically, the inspiration project focuses on the development of information retrieval and ranking methods for discovering relationships between RNA/DNA molecules, proteins, diseases, and anti-cancer drugs. The system proposed for the BioASQ Synergy 2026 task uses a large language model to extract biomedical keywords from user questions. The extracted keywords are then passed to Elasticsearch that retrieves candidate PubMed documents based on the BM25 information retrieval model with phrase boosting and all-keyword matching, followed by cross-encoder reranking of the retrieved results. The proposed system incorporates in the final query processing stage LLM-based reranking to further refine the ranking of top candidate documents.

Keywords

biomedical information retrieval, BioASQ Synergy, retrieval-augmented generation, large language models, Elasticsearch, BM25, cross-encoder reranking, biomedical question answering

1. Introduction

The rapid growth of biomedical literature has made it increasingly difficult for researchers and clinicians to identify relevant evidence efficiently. Large databases such as PubMed contain millions of scientific articles, and new publications are added continuously. As a result, biomedical question answering systems must not only retrieve relevant documents, but also rank evidence effectively and support answer generation from reliable sources.

Biomedical information retrieval is especially challenging because biomedical questions often contain specialized terminology, gene and protein names, disease names, drug names, abbreviations, and synonyms. A simple keyword-based search may fail when the query terms do not exactly match the terminology used in relevant articles. At the same time, large language models (LLMs) can generate fluent responses but may produce unsupported or hallucinated information when they are not grounded in retrieved evidence [1]. These challenges motivate retrieval-augmented approaches that combine the language understanding ability of LLMs with evidence retrieval from biomedical literature.

This paper presents an LLM-guided retrieval-augmented biomedical retrieval system developed for the BioASQ Synergy 2026 task. The system uses an LLM to extract biomedical keywords from each question. Because LLM-generated keywords may include overly broad terms, duplicated concepts, or terms that are not useful for retrieval, the extracted keywords are filtered using a custom-developed list of generic biomedical terms. The filtered keywords are then used to retrieve candidate PubMed documents with Elasticsearch BM25. To improve ranking quality, the system applies cross-encoder reranking to the retrieved candidates and integrates LLM-based reranking of intermediate results to refine the top-ranked documents.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding authors.

srattan2@depaul.edu (S. Rattanapani); dstan@cdm.depaul.edu (D. Raicu); jfurst@cdm.depaul.edu (J. Furst); aorhean@depaul.edu (A. I. Orhean)

The main contributions of this work are as follows:

- We adapt retrieval ideas from an ongoing unpublished genomic and proteomic search project to the BioASQ Synergy task, including biomedical concept extraction, Elasticsearch-based retrieval, synonym-aware search, and multi-stage ranking for evidence discovery.
- We develop an LLM-guided biomedical keyword extraction module that generates focused query terms from BioASQ questions while preserving important biomedical concepts.
- We filter LLM-extracted keywords using a custom-developed list of generic biomedical terms to reduce overly broad, duplicated, or low-information query terms before retrieval.
- We build a retrieval pipeline using Elasticsearch BM25 with phrase boosting and all-keyword matching over PubMed-style documents.
- We incorporate a two-stage reranking process, where a cross-encoder first reranks candidate documents retrieved by BM25, and an LLM then reranks the top 20 documents to select the final top 10 submissions.
- We evaluate the system using official BioASQ Synergy 2026 results and analyze its strengths and limitations across document retrieval, snippet retrieval, exact answers, and ideal answers.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the system architecture. Section 4 presents the main search stages. Section 5 describes the experimental setup. Section 6 reports the official BioASQ results. Section 7 concludes the paper.

2. Related Work

This work is inspired by an ongoing project that investigated literature-based discovery of aptamer–protein interactions using indexing, search, and large language models. In that project, we focused on identifying proteins associated with a specific drug–disease context, using the AS1411 aptamer and Kaposi's sarcoma-associated herpesvirus (KSHV) as a case study. The system used Elasticsearch indexing, synonym-aware protein normalization, proximity-based evidence retrieval, and LLM-based reasoning to retrieve and rank candidate proteins from biomedical literature.

The BioASQ Synergy system presented in this paper adapts several ideas from that inspiration project to a biomedical question answering setting. Instead of retrieving proteins for a fixed drug–disease or molecule–disease context, the system retrieves PubMed documents relevant to natural-language biomedical questions. The shared goal is to use scalable indexing, biomedical keyword extraction, and multi-stage ranking to support evidence discovery from biomedical literature.

Retrieval-augmented generation combines language models with external document retrieval so that generated outputs can be grounded in retrieved evidence [1]. This approach is useful for biomedical search because scientific answers should be supported by relevant literature rather than generated from model knowledge alone.

Lexical retrieval methods such as BM25 remain widely used for large-scale document retrieval because they are efficient and effective over large text collections. However, lexical retrieval may miss relevant documents when biomedical concepts are expressed using different synonyms, abbreviations, or related terminology. This limitation motivates the use of synonym-aware search and Elasticsearch-based indexing in the inspiration project, and it also motivates the BM25 candidate retrieval stage in this BioASQ system.

Neural reranking methods address some limitations of lexical retrieval by estimating semantic relevance between a query and a candidate document. Cross-encoder reranking models jointly encode the query and candidate text and can improve ranking quality after an initial retrieval stage. Hybrid retrieval methods further combine lexical and semantic relevance signals, often through normalized score fusion [2].

The BioASQ challenge provides a shared benchmark for biomedical semantic indexing and question answering. It evaluates systems on tasks such as document retrieval, snippet retrieval, exact-answer

generation, and ideal-answer generation. The Synergy task further emphasizes iterative biomedical question answering, where systems retrieve evidence and generate answers for developing biomedical questions. Official BioASQ overview papers provide the task and evaluation context for this work [3, 4, 5, 6].

Recent BioASQ participant systems also use retrieval-augmented and multi-stage retrieval architectures. One BioASQ@CLEF 2025 system uses a multi-stage RAG pipeline that combines question analysis, biomedical entity extraction, semantic retrieval, and LLM-based answer generation [7]. Another BioASQ 13b system combines BM25 retrieval over PubMed titles and abstracts, reranking, and LLM-based answer generation [8]. A third system explores dense retrieval, cross-encoder reranking, GPT-based reranking, and LLM prompting for biomedical question answering [9]. These systems show that recent BioASQ approaches increasingly combine lexical retrieval, neural reranking, and LLM-based reasoning.

Our system follows this general direction but differs by using an LLM-guided biomedical keyword extraction step together with a custom-developed generic biomedical term filter before Elasticsearch BM25 retrieval, followed by weighted BM25–cross-encoder score fusion and LLM-based final reranking.

3. System Overview

Figures 1 and 2 show the overall design of the proposed LLM-guided retrieval-augmented biomedical retrieval system. The system contains two main parts: an offline indexing stage and an online retrieval and reranking stage.

Figure 1 illustrates the offline indexing stage. The biomedical document collection is first collected from PubMed. Each article is represented by three fields: PubMed identifier, title, and abstract. The PubMed identifier is stored as a `keyword` field so that document identifiers can be preserved exactly for lookup and BioASQ submission. The title and abstract are stored as `text` fields using the custom `index_analyzer`. These two fields are used for BM25 keyword matching, phrase matching, all-keyword matching, cross-encoder reranking input, snippet extraction, and answer-generation evidence. This indexed table allows the system to efficiently search over PubMed-style documents during the BioASQ Synergy task. In this work, the retrieval system focuses mainly on the title and abstract fields because BioASQ Synergy requires relevant snippets to be selected from the title or abstract of returned articles.

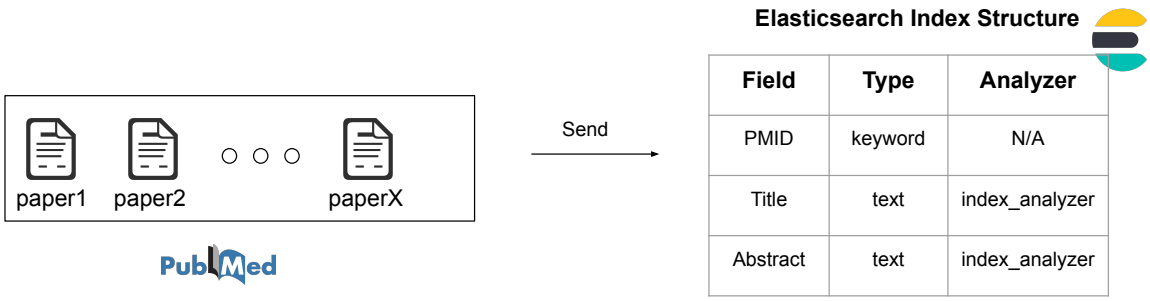

Figure 1: Offline indexing stage. PubMed articles are represented by PMID, title, and abstract fields and stored in an Elasticsearch index for later retrieval.

Figure 2 shows the online retrieval and reranking pipeline used for each BioASQ question. Given a biomedical question, the system first uses a large language model (LLM) to extract important biomedical keywords or phrases. These keywords represent the main concepts in the question, such as diseases, genes, proteins, drugs, or biological processes. The extracted keywords are then used to search the Elasticsearch index.

The first retrieval stage uses Elasticsearch BM25 to retrieve a large candidate set of documents. In our system, BM25 returns up to 10,000 candidate papers for each question. This stage is designed to

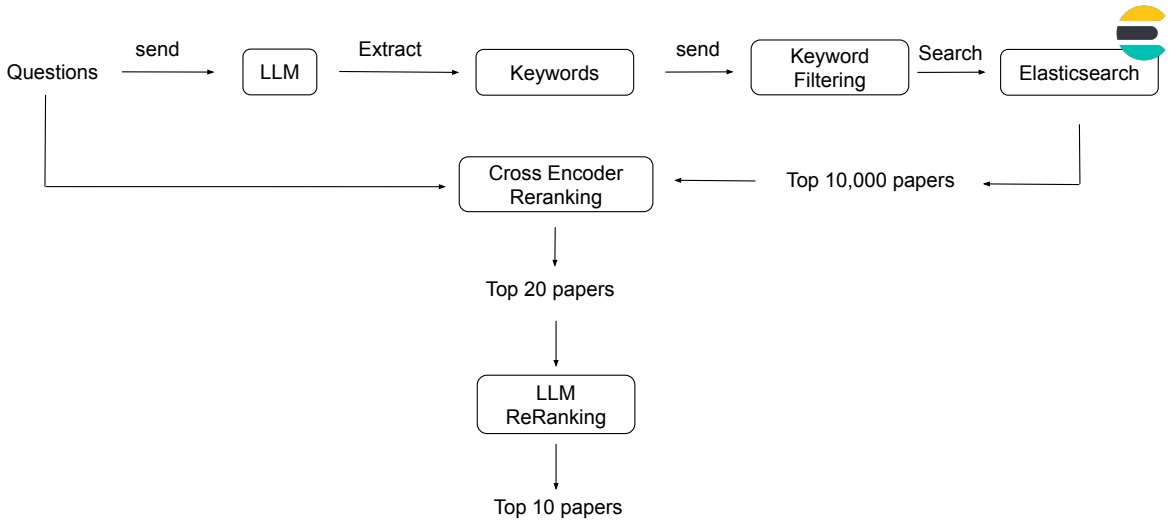

Figure 2: Online retrieval and reranking pipeline. The system extracts keywords from a BioASQ question, retrieves the top 10,000 candidate papers using Elasticsearch BM25, reranks them with a cross-encoder, and uses an LLM to select the final top 10 papers.

maximize recall by collecting a broad set of potentially relevant PubMed articles. The BM25 query combines keyword matching, phrase boosting, and all-keyword matching so that documents containing important biomedical terms are ranked higher.

After BM25 retrieval, the system applies a cross-encoder reranking model. The cross-encoder compares the original BioASQ question with each candidate document and assigns a relevance score. Prior work on hybrid retrieval has shown that lexical and semantic relevance signals can be combined using normalized weighted score fusion [2]. Following this idea, the system normalizes the BM25 and cross-encoder scores before combining them. The final reranking score is computed as:

$$S(d) = (1 - w) \cdot S_{CE}(d) + w \cdot S_{BM25}(d) \tag{1}$$

where $S_{CE}(d)$ is the normalized cross-encoder score for document d, $S_{BM25}(d)$ is the normalized BM25 score, and w controls the contribution of the BM25 score. In our implementation, $w = 0.15$, giving greater importance to the cross-encoder score while retaining a smaller contribution from BM25. This weighted fusion allows the ranking to preserve lexical relevance from BM25 while incorporating semantic relevance from the cross-encoder. The value $w = 0.15$ was selected heuristically based on the intuition that the cross-encoder should dominate the final relevance estimate after candidate retrieval, while BM25 should still contribute a small lexical matching signal. Systematic tuning of w across development queries is left for future work.

After this reranking stage, the top 20 documents are passed to an LLM-based reranker. The LLM is prompted to judge which documents are most useful for answering the biomedical question based on the available title and abstract information. The LLM reranks these top 20 documents, and the final output is a ranked list of the top 10 PubMed articles formatted as a BioASQ submission file.

Overall, the system follows a multi-stage retrieval design: PubMed indexing, LLM-guided keyword extraction, BM25 candidate retrieval, BM25-weighted cross-encoder reranking, and LLM-based final reranking. This design allows the system to combine efficient large-scale lexical retrieval with semantic reranking and LLM-based relevance judgment.

4. Search Stages

This section describes the main components of the proposed system in more detail. The system consists of five main components: LLM-guided keyword extraction, BM25 candidate retrieval, cross-encoder

reranking, LLM-based final reranking, and BioASQ submission generation.

4.1. LLM-Guided Keyword Extraction

For each BioASQ question, the system first uses an LLM to extract a small set of biomedical keywords or phrases. The goal of this step is to transform a natural-language biomedical question into focused search terms that can be used by the retrieval engine. These keywords are intended to capture the main biomedical concepts in the question, such as diseases, genes, proteins, drugs, abbreviations, or biological processes.

Because LLM-generated keywords may include overly broad terms, duplicated concepts, or terms that are not useful for retrieval, the extracted keywords are filtered using a custom-developed list of generic biomedical terms. Examples of generic terms include broad words such as *disease*, *protein*, *gene*, *treatment*, *patient*, and *cell*, which may not be useful as standalone search terms. This filtering step also removes empty terms and duplicate keywords. In addition, the system keeps only keywords that appear in the original question, which helps reduce the risk of introducing unsupported or hallucinated query terms.

The final filtered keywords are then used as input to the BM25 retrieval stage. By combining LLM-based keyword extraction with keyword filtering based on a custom-developed biomedical term list, the system aims to improve query focus while keeping the retrieved terms faithful to the original BioASQ question.

4.2. BM25 Candidate Retrieval

After keyword extraction, the filtered keywords are submitted to an Elasticsearch index containing PubMed-style documents. Each indexed document contains a PubMed identifier, title, and abstract. The system searches over the title and abstract fields because BioASQ Synergy document and snippet evidence is based on these fields.

The title and abstract fields were indexed as text fields using a custom Elasticsearch analyzer named `index_analyzer`. This analyzer uses the standard tokenizer followed by a lowercase token filter. The same analyzer was used at query time for keyword matching, phrase matching, and all-keyword matching. PubMed identifiers were stored as keyword fields so that returned PMIDs could be preserved exactly. The index used one shard and zero replicas, and Elasticsearch's default BM25 similarity was used for scoring.

The BM25 retrieval stage is designed to retrieve a large candidate set of potentially relevant documents. In our implementation, the system retrieves up to 10,000 candidate papers for each question. This large candidate set is intended to improve recall by increasing the chance that relevant documents are included before applying more expensive reranking models.

The Elasticsearch query combines several matching strategies over the filtered keyword list. First, the system applies standard keyword matching over the title and abstract fields. Second, it applies phrase matching with additional boosting so that documents containing exact biomedical phrases receive higher scores. Third, it includes an all-keyword matching condition with an additional boost, giving higher scores to documents that match all filtered query terms. Each retrieved candidate is stored with its PMID, title, abstract, and BM25 score.

For example, for the question "Are there any nucleosomes detected in the human serum?", if the LLM produces the terms *nucleosomes*, *human serum*, and *human*, the filtering step removes the standalone generic term *human* while retaining *human serum*. The final keyword list becomes *nucleosomes* and *human serum*. These filtered terms are then used in keyword-matching clauses over the title and abstract fields, with an additional boosted phrase-matching clause for the multi-word phrase *human serum* and an all-keyword matching clause over the filtered keyword list.

Although BM25 already reduces the influence of frequent terms through inverse document frequency, broad biomedical or question-related terms may still affect retrieval through query construction, especially when they appear in boosted phrase-matching clauses or all-keyword matching constraints.

The custom stoplist of generic biomedical and question-related terms is therefore used to reduce low-information LLM-generated terms before query construction. In our implementation, the post-filtering step removes empty terms, duplicate terms, terms containing non-English or unsupported characters, terms that do not appear in the original question, and terms that are either exact matches to generic banned words or consist entirely of banned tokens. We did not perform a separate ablation study isolating the effect of the stoplist in the current submission; a systematic comparison with and without the stoplist is left for future work.

4.3. Cross-Encoder Reranking

The candidate documents retrieved by BM25 are then reranked using the ms-marco-MiniLM-L-12-v2 cross-encoder model. Unlike BM25, which mainly relies on lexical matching, the cross-encoder jointly considers the BioASQ question and the candidate document text [10]. For each candidate document, the input to the cross-encoder is formed by pairing the original question with the document title and abstract. The model then produces a relevance score indicating how useful the document is likely to be for answering the question.

Because BM25 scores and cross-encoder scores are produced on different scales, the system normalizes both scores before combining them. Cross-encoder scores are normalized using min-max normalization:

$$S_{CE}(d_i) = \frac{s_{CE}(d_i) - \min_j s_{CE}(d_j)}{\max_j s_{CE}(d_j) - \min_j s_{CE}(d_j)} \tag{2}$$

where $s_{CE}(d_i)$ is the raw cross-encoder score for document d_i. BM25 scores are normalized by dividing each BM25 score by the maximum BM25 score among the candidate documents:

$$S_{BM25}(d_i) = \frac{s_{BM25}(d_i)}{\max_j s_{BM25}(d_j)} \tag{3}$$

The final reranking score is then computed as a weighted linear combination:

$$S(d) = (1 - w) \cdot S_{CE}(d) + w \cdot S_{BM25}(d) \tag{4}$$

where $S_{CE}(d)$ is the normalized cross-encoder score, $S_{BM25}(d)$ is the normalized BM25 score, and w controls the BM25 contribution. In our implementation, $w = 0.15$. This setting gives the cross-encoder the dominant contribution while retaining a smaller lexical matching signal from BM25. The value was selected heuristically as a small BM25 retention weight, and systematic tuning of this parameter is left for future work.

After the final score is computed, candidate documents are sorted in descending order by their combined score.

4.4. LLM-Based Final Reranking

After cross-encoder reranking, the system applies an LLM-based final reranking stage. This step is applied only to the top-ranked candidates after cross-encoder reranking in order to reduce computational cost. In our system, the top 20 documents are passed to the LLM-based reranker.

The LLM receives the original BioASQ question together with the PMID, title, and abstract of each candidate document. It is prompted to judge which documents are most useful for answering the biomedical question and to return a ranked list. This step allows the system to apply an additional relevance judgment based on the full question and the available document text.

The LLM-based reranker used the following prompt structure. The system message instructed the model to act as a biomedical information retrieval judge and to return only valid JSON. The user message contained the BioASQ question and the candidate PubMed documents. Each candidate document was represented by its rank index, PMID, title, and abstract. The prompt asked the model to rerank all candidate documents by usefulness for answering the biomedical question.

The prompt included the following instructions:

- Rerank the candidate documents by their usefulness for answering the biomedical question.
- Use the provided question, PMID, title, and abstract only.
- Return a JSON object containing a ranked list of documents.
- Include every candidate document exactly once.
- Do not include extra keys, explanations, or commentary.

The LLM-based reranker is used as the final refinement stage. After this step, the system selects the top 10 documents as the final document retrieval output for the BioASQ submission.

4.5. Submission Generation

The final component converts the system output into the BioASQ Synergy submission format. For each question, the system outputs the question identifier, question body, question type, and a ranked list of up to 10 PubMed document identifiers.

For answer generation, PubMed records retrieved from the Elasticsearch index were used as evidence. Each evidence record contained the PMID, title, and abstract. These records were provided to `gpt-4.1-mini`, which generated both an ideal answer and, when applicable, an exact answer. The ideal answer was formatted as a single text response.

Exact answers were formatted according to the BioASQ question type. Yes/no questions were normalized to `yes` or `no`, factoid questions were limited to up to five concise answers, and list questions were formatted as a list of answer items. Summary questions did not require an exact answer.

Snippet extraction was not implemented as a separate optimized passage-selection module in the submitted system. When snippets were included, they were selected from the title or abstract text of the retrieved PubMed documents, following the BioASQ Synergy requirement that snippets come from returned articles rather than from generated text. When snippet or answer outputs were not produced, the corresponding submission fields were left empty. The submitted system therefore focused primarily on document-level retrieval, with answer generation included for selected answer-ready questions.

5. Experimental Setup

5.1. Dataset and Task Setting

The system was developed and evaluated for the BioASQ Synergy 2026 task. The task evaluates biomedical retrieval and question answering systems across multiple test rounds. For each question, participating systems submit a ranked list of relevant PubMed documents and, when applicable, snippets, exact answers, and ideal answers.

Before participating in the official BioASQ Synergy 2026 evaluation, we used the BioASQ13 dataset as a development set to test different system configurations. This development evaluation was used mainly to compare different LLM-based reranking models and to select the final configuration used for official submission. The system was evaluated using document-level retrieval performance, with mean average precision (MAP) used as the primary selection metric during development.

For the official BioASQ Synergy 2026 submission, our system was submitted under the name *AgenticAI and RAG*. Although the submitted system supported several output fields in the BioASQ format, the main focus of the system was document retrieval. Therefore, the strongest and most consistent evaluation results were expected in the document retrieval task.

5.2. System Configuration

The final system configuration used a multi-stage retrieval and reranking pipeline. First, LLM-generated biomedical keywords were used to retrieve candidate documents from an Elasticsearch index using BM25. The system retrieved up to 10,000 candidate documents for each question. These candidates were then reranked using the `cross-encoder/ms-marco-MiniLM-L-12-v2` cross-encoder model.

Table 1
Development evaluation of LLM-based final reranking models on the BioASQ13 dataset. MAP was used as the model-selection metric.

Model	Organization	MAP
`gpt-4.1-mini`	OpenAI	0.521
`gpt-oss-120b`	OpenAI	0.497
`gpt-oss-20b`	OpenAI	0.480
`Meta-Llama-3.1-8B-Instruct`	Meta	0.450
`Meta-Llama-3.1-70B-Instruct`	Meta	0.491
`Meta-Llama-3.1-405B-Instruct`	Meta	0.494
`Llama-3.3-70B-Instruct`	Meta	0.491
`Llama-4-Scout-17B-16E-Instruct`	Meta	0.488
`gemma-3-27b-it`	Google	0.469

After cross-encoder reranking, the top 20 documents were passed to an LLM-based reranker, and the final top 10 documents were selected for submission.

During development, we evaluated several LLMs for the final reranking stage using the BioASQ13 dataset. The purpose of this experiment was to select the LLM used in the final reranking step before submitting to the official BioASQ Synergy 2026 evaluation. Most models were accessed through the Argonne National Laboratory LLM inference service, while `gpt-4.1-mini` was accessed through the OpenAI API. MAP was used as the model-selection metric because the system focused mainly on ranked document retrieval.

As shown in Table 1, `gpt-4.1-mini` achieved the highest development MAP score of 0.521. Based on this result, we selected `gpt-4.1-mini` as the final LLM reranker for the official BioASQ Synergy 2026 submission.

The final configuration can be summarized as follows:

- BM25 candidate retrieval with up to 10,000 candidate documents;
- cross-encoder reranking using `cross-encoder/ms-marco-MiniLM-L-12-v2`;
- weighted score fusion between normalized BM25 and cross-encoder scores;
- LLM-based reranking of the top 20 documents using `gpt-4.1-mini`;
- final submission of the top 10 ranked PubMed documents.

6. Results

Table 2 summarizes the official BioASQ Synergy 2026 document retrieval results for our submitted system, listed as *AgenticAI and RAG* in the official leaderboard. Overall, the system achieved its strongest results in document retrieval and selected ideal-answer settings. In Test Round 1, the system achieved the highest mean precision, F-measure, and MAP among participating systems for document retrieval. In Test Round 2, the system achieved the highest MAP for document retrieval and also obtained the best automatic ROUGE-based scores for ideal-answer generation. Official results also show that the system achieved strong manual ideal-answer scores in Test Round 2, including readability, recall, precision, and repetition.

In Test Round 1 document retrieval, the system obtained a mean precision of 0.5295, recall of 0.4281, F-measure of 0.4235, and MAP of 0.5378. These scores ranked the system first among participating systems in mean precision, F-measure, and MAP for document retrieval in that round. In Test Round 2 document retrieval, the system obtained a MAP of 0.3370, which was the highest MAP among the listed systems for that round. In Test Round 4, the system achieved the highest document F-measure, with a score of 0.2882. These results indicate that the proposed retrieval pipeline was effective at ranking relevant PubMed documents in several official evaluation settings.

Table 2
Official BioASQ Synergy Task 2026 document retrieval results for the submitted system *AgenticAI and RAG*. Values are reported as score followed by the official rank in parentheses among systems with valid scores for that metric.

Round	Mean Precision	Recall	F-Measure	MAP	GMAP
Round 1 Documents	0.5295 (1)	0.4281 (4)	0.4235 (1)	0.5378 (1)	0.3293 (4)
Round 2 Documents	0.3320 (1)	0.3743 (4)	0.3018 (4)	0.3370 (1)	0.0387 (4)
Round 3 Documents	0.2578 (4)	0.3352 (3)	0.2573 (4)	0.2318 (4)	0.0304 (5)
Round 4 Documents	0.2444 (1)	0.4608 (2)	0.2882 (1)	0.2905 (4)	0.0403 (4)

Table 3
Official BioASQ Synergy 2026 Test Round 2 ideal-answer results for the submitted system.

Metric	R-2 Rec.	R-2 F1	R-SU4 Rec.	R-SU4 F1
Automatic score	0.4071	0.2826	0.4150	0.2849

Metric	Readability	Recall	Precision	Repetition
Manual score	4.60	4.55	4.29	4.67

Table 3 shows the Test Round 2 ideal-answer results. This was one of the strongest outcomes of our system. The submitted system achieved the highest automatic ROUGE-based scores among participating systems, including R-2 recall, R-2 F1, R-SU4 recall, and R-SU4 F1. The system also received strong manual evaluation scores, with readability of 4.60, recall of 4.55, precision of 4.29, and repetition of 4.67.

These results demonstrate that the proposed pipeline was effective not only for retrieving relevant biomedical documents, but also for supporting answer generation from retrieved evidence. Ideal-answer generation requires the system to identify useful biomedical evidence and generate a coherent natural-language response based on that evidence. Therefore, the strong ideal-answer performance provides evidence that the retrieved documents were relevant and informative enough for the LLM to generate high-quality biomedical answers. This result highlights the value of combining LLM-guided retrieval, cross-encoder reranking, and LLM-based generation in a retrieval-augmented biomedical question answering system.

Although performance varied across subtasks and rounds, the system showed clear strengths in document retrieval and ideal-answer generation. Snippet retrieval was weaker than document retrieval, especially in later rounds, indicating that passage-level evidence selection remains an important area for improvement. Exact-answer performance was also incomplete in some rounds, particularly for factoid and list questions. However, the Test Round 2 ideal-answer results were a major strength of the system. The system achieved the highest automatic ROUGE-based scores among participating systems and received strong manual scores for readability, recall, precision, and repetition. These results suggest that the retrieved and reranked documents provided high-quality biomedical evidence for LLM-based answer generation. Overall, the official results indicate that the proposed LLM-guided retrieval and reranking pipeline was effective not only for identifying relevant biomedical articles, but also for supporting evidence-grounded biomedical answer generation.

7. Conclusion

This paper presented an LLM-guided retrieval-augmented biomedical retrieval system developed for the BioASQ Synergy 2026 task. The system combines LLM-based keyword extraction, custom-developed keyword filtering, Elasticsearch BM25 candidate retrieval, cross-encoder reranking, weighted score fusion, and LLM-based final reranking. The final system retrieves up to 10,000 candidate PubMed documents, reranks them using a cross-encoder, and then applies an LLM reranker to select the final

top 10 documents for submission.

Development experiments on the BioASQ13 dataset showed that the choice of LLM for final reranking affected retrieval performance. Among the tested models, `gpt-4.1-mini` achieved the highest MAP score and was selected as the final LLM reranker for the official BioASQ Synergy 2026 submission.

Official BioASQ Synergy 2026 results showed that the system achieved strong performance in document retrieval and ideal-answer generation. In Test Round 1, the system achieved the highest mean precision, F-measure, and MAP among participating systems for document retrieval. In Test Round 2, the system achieved the highest automatic ROUGE-based ideal-answer scores and strong manual evaluation scores. These results suggest that the retrieval pipeline was able to identify useful biomedical evidence and support high-quality LLM-based answer generation.

However, the system was less consistent for snippet retrieval and structured exact-answer generation. This limitation is partly due to the system's main focus on document-level retrieval rather than fine-grained passage extraction or answer-type-specific generation. Future work should focus on dedicated snippet extraction, answer-type-specific modules for yes/no, factoid, and list questions, entity normalization, and systematic tuning of reranking parameters. Future work should also include an ablation study measuring the isolated effect of the custom biomedical term stoplist, since BM25 already downweights frequent terms through inverse document frequency and the stoplist may interact differently with keyword matching, phrase boosting, and all-keyword matching. In future versions, the system can be extended from a document-focused retrieval pipeline into a more complete biomedical question answering system that jointly optimizes document retrieval, snippet selection, exact answers, and ideal-answer generation.

Acknowledgments

The authors gratefully acknowledge Argonne National Laboratory and NSF Chameleon Cloud for providing computational resources used during the development and evaluation of this work. The authors also thank the system administrators and infrastructure providers for supporting the experimental environment.

This work was supported in part by project resources from DePaul University and Rosalind Franklin University of Medicine and Science.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT web for grammar checking, wording suggestions, and sentence polishing. The author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

The system described in this paper also uses large language models as part of the research method, including LLM-guided keyword extraction, LLM-based reranking, and LLM-based answer generation. These uses are described in the system methodology and are part of the proposed biomedical retrieval and question answering system.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[2] S. Bruch, S. Gai, A. Ingber, An analysis of fusion functions for hybrid retrieval, ACM Transactions on Information Systems 42 (2023) 1–35.

[3] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, et al., Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering,

in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 173–198.

[4] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC bioinformatics 16 (2015) 138.

[5] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[6] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 14b and Synergy14 in CLEF2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[7] S. Dueñas-Romero, L. A. U. López, E. Martínez-Cámara, Sinai at bioasq@clef 2025: A multi-stage rag pipeline for biomedical semantic question answering, in: Conference and Labs of the Evaluation Forum, 2025. URL: https://api.semanticscholar.org/CorpusID:282150093.

[8] J.-C. Han, B.-C. Chih, H.-C. Hung, R. T.-H. Tsai, Ncu-iisr: A retrieval-augmented generation approach for bioasq 13b phase a and a+, in: Conference and Labs of the Evaluation Forum, 2025. URL: https://api.semanticscholar.org/CorpusID:282324706.

[9] S. Verma, F. Jiang, X. Xue, Beyond retrieval: ensembling cross-encoders and gpt rerankers with llms for biomedical qa, arXiv preprint arXiv:2507.05577 (2025).

[10] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv preprint arXiv:1901.04085 (2019).