MedCascade: A Multi-Stage Cascaded Retrieval and Generation System for Biomedical Question Answering*

BioASQ at CLEF 2026

Artem Avdeiko[1,2], Pelinsu Topaloglu[1,2], Teresa Vaccari[1,2] and Georgios Peikos[2,*]

[1]University of Pavia, 27100 Pavia, Italy
[2]University of Milano-Bicocca (DISCo), 20126 Milan, Italy

Abstract

This paper presents MedCascade, our submission to BioASQ task 14B for biomedical question answering. The task evaluates systems across document and snippet retrieval, exact answer generation, and ideal answer generation. MedCascade combines LLM-based query expansion, sparse retrieval, neural reranking, LLM-based listwise reranking, Dense PRF reranking, snippet extraction, and LLM-based answer generation. Across the submitted batches, MedCascade-v1 achieved the strongest performance among our submitted systems. In Phase A, it ranked 9th and 4th for document retrieval in batches 3 and 4, respectively, and 8th and 7th for snippet retrieval. In Phase A+, it ranked 3rd for both exact and ideal answers in batch 3, and 1st for ideal answers in batch 4. The results suggest that high-quality evidence retrieval is important for biomedical QA, while exact answer generation remains sensitive to question type and output format.

Keywords

Biomedical question answering, BioASQ, Information retrieval, Retrieval augmented generation, Neural reranking, Dense pseudo relevance feedback

1. Introduction

Biomedical question answering is a central task in biomedical information access, requiring systems to retrieve, select, and synthesize evidence from large collections of scientific literature in response to expert-formulated information needs. In benchmark settings such as BioASQ [1, 2], this task evaluates not only the ability of systems to identify relevant articles and snippets, but also their capacity to generate exact answers and concise ideal answers that are faithful to the supporting biomedical evidence.

Recent work has shown that large language models (LLMs) have become central components in biomedical question answering systems, in line with their strong performance across medical, biomedical, and clinical NLP tasks [3, 4]. Current biomedical QA systems, however, are not limited to direct LLM-based answer generation. They include prompted LLMs, fine-tuned models, and hybrid pipelines that combine evidence retrieval, reranking, and answer generation [3, 5, 6]. Recent systems also use model adaptation methods, such as parameter-efficient fine-tuning, to improve the use of biomedical evidence and the quality of generated answers [7]. Retrieval augmented generation (RAG) has become an important direction in this space, as it allows systems to retrieve relevant biomedical evidence before generating answers [7]. This direction is also reflected in recent systems submitted to the BioASQ evaluation labs, which combine traditional information retrieval methods with LLMs, leverage query expansion, use knowledge resources, and perform RAG-based generation [8]. Nevertheless, recent BioASQ results show that access to relevant evidence substantially improves answer quality, suggesting that retrieval quality remains a key factor for the effectiveness of RAG-based biomedical QA systems [1].

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
artem.avdeiko01@universitadipavia.it (A. Avdeiko); pelinsu.topaloglu01@universitadipavia.it (P. Topaloglu); teresa.vaccari01@universitadipavia.it (T. Vaccari); georgios.peikos@unimib.it (G. Peikos)
0009-0004-7448-9272 (A. Avdeiko); 0000-0002-0877-7063 (P. Topaloglu); 0000-0001-7116-9338 (T. Vaccari); 0000-0002-2862-8209 (G. Peikos)

This is particularly important because RAG pipelines may underperform direct LLM-based generation when the retrieved context is not sufficiently relevant or useful for answering the question [9].

The BioASQ task 14B provides a benchmark setting for evaluating the connection between evidence retrieval and biomedical answer generation across three phases. In Phase A, systems retrieve relevant PubMed documents and snippets for biomedical questions in English, in Phase A+ they additionally provide exact and ideal answers, and in Phase B they generate updated exact and ideal answers using expert-selected relevant material.

In this work, we present MedCascade, a biomedical QA system based on a cascaded retrieval and generation pipeline. Our main design goal was to combine high-recall first-stage retrieval with successive reranking stages, aiming to progressively improve precision in the final top-10 documents used for answer generation. To this aim, MedCascade combines LLM-based query expansion, lexical retrieval, semantic reranking, LLM-based listwise reranking, and a dense PRF reranking stage to improve the quality of the retrieved evidence. MedCascade is an effectiveness-oriented biomedical QA system that prioritizes retrieval quality over latency, with the most computationally demanding reranking stages parallelized across multiple GPUs to reduce execution time. Overall, MedCascade is best viewed as an effectiveness-oriented integration of established retrieval and reranking components, empirically validated in a biomedical QA setting.

We participated in BioASQ task 14B [10] test batches 2, 3, and 4, submitting several systems and MedCascade variants that differed in their retrieval, reranking, and generation components. Our submissions are summarized in Table 1, while the methodology section describes the MedCascade configuration that achieved the strongest overall performance among our submissions. The results section then reports the performance of the submitted systems across the relevant BioASQ phases and test batches.

2. Related Work

MedCascade builds on prior work in biomedical QA that combines lexical retrieval, neural reranking, LLM-based query expansion, PRF, and LLM-based answer generation. This section reviews these directions, focusing on retrieval, reranking, query reformulation, and retrieval-augmented answer generation in biomedical QA.

2.1. Biomedical Evidence Retrieval

First-stage retrieval in biomedical QA systems is commonly based on sparse, dense, or hybrid retrieval approaches [11, 12, 13]. Although neural retrieval methods have become increasingly effective [14], BM25 [15] remains a strong and widely used model in BioASQ systems, both as a standalone first-stage retriever and as part of more complex multi-stage pipelines [12, 13]. Its efficiency and robustness make it particularly suitable for retrieving large candidate sets that can then be refined by more computationally expensive reranking models.

At the same time, biomedical evidence retrieval has increasingly moved beyond purely lexical matching. Neural sparse methods, such as SPLADE, preserve sparse representations while introducing semantic expansion learned from language models, offering an alternative to classical term matching [16]. Dense retrieval methods, often implemented through bi-encoders or sentence embedding models, are also used to retrieve documents based on learned semantic representations of questions and documents [17, 18]. These approaches complement lexical retrieval by matching questions and documents in a learned semantic space, although their effectiveness depends on domain adaptation and representation quality.

For this reason, several BioASQ systems adopt hybrid retrieval strategies that combine sparse and dense signals [13, 19, 20, 11]. A common strategy is to combine BM25 with BERT-based bi-encoders, using lexical matching and semantic similarity as complementary evidence signals [13, 21]. In practice, hybrid retrieval often requires separate sparse and dense indexing pipelines, motivating recent work on joint learning approaches that aim to integrate these signals more directly [20].

2.2. Neural and LLM-based Reranking

Reranking is commonly used in biomedical QA systems to refine the candidate set produced by first-stage retrieval. More expressive models, such as cross-encoders, are typically applied to a smaller subset of documents in order to improve early precision before snippet selection or answer generation [22, 23, 24, 25, 13]. Cross-encoders have shown strong performance in BioASQ systems and have been used in multi-stage pipelines that combine reranking with dense PRF, although their computational cost limits their use over very large candidate sets [18, 23, 13].

Recent work has also explored LLM-based reranking, with pointwise and listwise formulations being the most common. Pointwise approaches score candidate documents independently, for example by estimating the likelihood of a relevance-indicating response [26, 27]. Listwise approaches instead ask the model to reorder a list of candidate documents, allowing the reranker to compare documents jointly [28, 5, 6]. Prior work has shown that listwise reranking can improve over pointwise baselines, although its applicability is constrained by context length and inference cost [29, 30]. In BioASQ, LLM-based reranking has also been explored, with Verma et al. [18] reporting improvements from a GPT-4o listwise reranker. These results motivate the use of LLM-based listwise reranking as a high-precision component applied after earlier retrieval and reranking stages.

2.3. Query Expansion and PRF

Query expansion and PRF are established query reformulation strategies for addressing vocabulary mismatch, which is especially important in biomedical retrieval because relevant evidence may be expressed through synonyms, abbreviations, ontology terms, among others [31, 32].

Query expansion has long been used to reformulate user queries and improve retrieval effectiveness [33]. Recent query expansion methods increasingly use LLMs to generate expansion terms, synthetic snippets, or ontology-guided reformulations [34, 35, 36, 37, 38]. These methods can introduce relevant biomedical terminology that is absent from the original question, thereby improving the match between the query and target evidence. At the same time, generated expansions require careful control, since they may introduce noise or shift retrieval away from the original information need.

PRF is a standard technique for improving first-stage retrieval by using initially retrieved documents to refine the query representation [39, 40]. However, its effectiveness depends on the quality of the documents selected as pseudo relevant, since noisy feedback documents can shift the query representation away from the original information need [41]. Recent biomedical QA systems have adapted PRF to both sparse and dense retrieval settings. In BioASQ, Borazio et al. [37] adopted a snippet-based PRF strategy in which high-confidence snippets from the initially retrieved documents were used as feedback terms to reformulate the BM25 query and refine the candidate ranking. Dense PRF methods, including ColBERT-PRF and related dense feedback approaches, follow a different strategy by using the vector representations of pseudo relevant documents to refine the query representation or adjust document scoring in the embedding space [42, 43, 44, 23]. In this setting, feedback documents are not used as additional query text, but as semantic signals that guide the ranking model toward documents that are closer to the initially identified relevant evidence.

2.4. Answer Generation in Biomedical QA

RAG has become a common strategy for the answer generation phases of BioASQ, where systems use retrieved documents or snippets as context for generating exact and ideal answers. Recent systems have explored different ways of selecting and presenting this evidence to LLMs, including snippet selection, task-aware prompting, and model ensembling [37, 45, 46, 47]. For example, Kim et al. [45] show that prompting strategies should be adapted to the question type, since different types of questions may benefit from different ways of presenting snippets to the model. Ensemble approaches based on multiple proprietary LLMs have also achieved competitive performance, especially for yes/no questions [46]. These studies suggest that answer generation in biomedical QA depends not only on the choice of the LLM, but also on how retrieved evidence is selected, structured, and provided to the model.

Overall, prior work shows that biomedical QA systems increasingly rely on multi-stage architectures that combine retrieval, reranking, query expansion, PRF, and LLM-based answer generation. MedCascade follows this direction, focusing on the progressive refinement of retrieved evidence before answer generation.

3. MedCascade System Overview

MedCascade is a multi-stage biomedical QA system designed to balance high-recall first-stage retrieval with high-precision evidence refinement through successive reranking stages.[1] The system starts from a large candidate set retrieved from the indexed PubMed corpus[2] and progressively applies more selective reranking stages to improve the precision of the final ranked list. The pipeline combines LLM-based query expansion, BM25 first-stage retrieval, intermediate neural reranking, LLM-based listwise reranking, Dense PRF reranking, score fusion, snippet extraction, and LLM-based answer generation. The main retrieval objective is to improve the quality of the top-ranked documents used for snippet extraction and answer generation. In the final retrieval stage, MedCascade operates on the top 1000 candidates, applies LLM-based listwise reranking to the current top 20 documents, uses the top $k = 3$ documents as pseudo relevant feedback, and then re-scores the full top-1000 candidate set. Figure 1 provides an overview of the retrieval and reranking pipeline.

3.1. First-stage Retrieval

Given a biomedical question q, MedCascade first applies LLM-based query expansion to generate a set of keywords, synonyms, and related terms, denoted as t'. We use Claude Sonnet 4[3] in a zero-shot setting to generate these expansion terms. The expansion terms are then combined with the original question to form an expanded query.

Unless otherwise stated, all preliminary experiments reported in this section were conducted on BioASQ task 13B 2025 test batch 4[4]. In these experiments, we observed that simply concatenating the original question with the generated expansion terms reduced retrieval performance, particularly recall. To mitigate this effect, we followed prior work that repeats the original query before appending generated expansions [48, 49]. The intuition is to preserve the weight of the original information need while still allowing the retrieval model to benefit from additional related terms. The expanded query is defined as:

$$q^{+} = \mathrm{concat}\left(q^{\times R}, t'\right)$$

where $q^{\times R}$ denotes the repetition of the original query R times. In MedCascade, we set $R = 15$, as our preliminary experiments showed a performance plateau followed by a slow decrease.

The document corpus used for retrieval is the PubMed Baseline Repository 2026[2]. This corpus contains citations and abstracts from biomedical literature indexed by NLM resources, including MEDLINE, and corresponds to the document source used in BioASQ task 14B. The expanded query q^{+} is used to retrieve candidate documents from this indexed PubMed corpus. We use BM25 as the first-stage retriever because it provides an efficient and robust mechanism for retrieving a large initial candidate set of documents. To mitigate noise introduced by query expansion, we also perform retrieval with the original query q. The rankings obtained from q and q^{+} are then fused using CombSUM [50], producing the final first-stage ranking by summing the corresponding BM25 scores. We retain the top 3000 documents from this ranking in order to obtain a high-recall candidate set for the subsequent reranking stages.

[1] The implementation of MedCascade, together with supplementary material, analyses and future developments, is available at https://github.com/mitruhaa/bioasq_systems.
[2] Indexed PubMed corpus, accessed on 2/7/26: https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
[3] Claude Sonnet 4 model, accessed on 2/7/26: https://www.anthropic.com/claude/sonnet
[4] Data used in our preliminary experiments, accessed on 2/7/26: https://participants-area.bioasq.org/Tasks/b/testData/

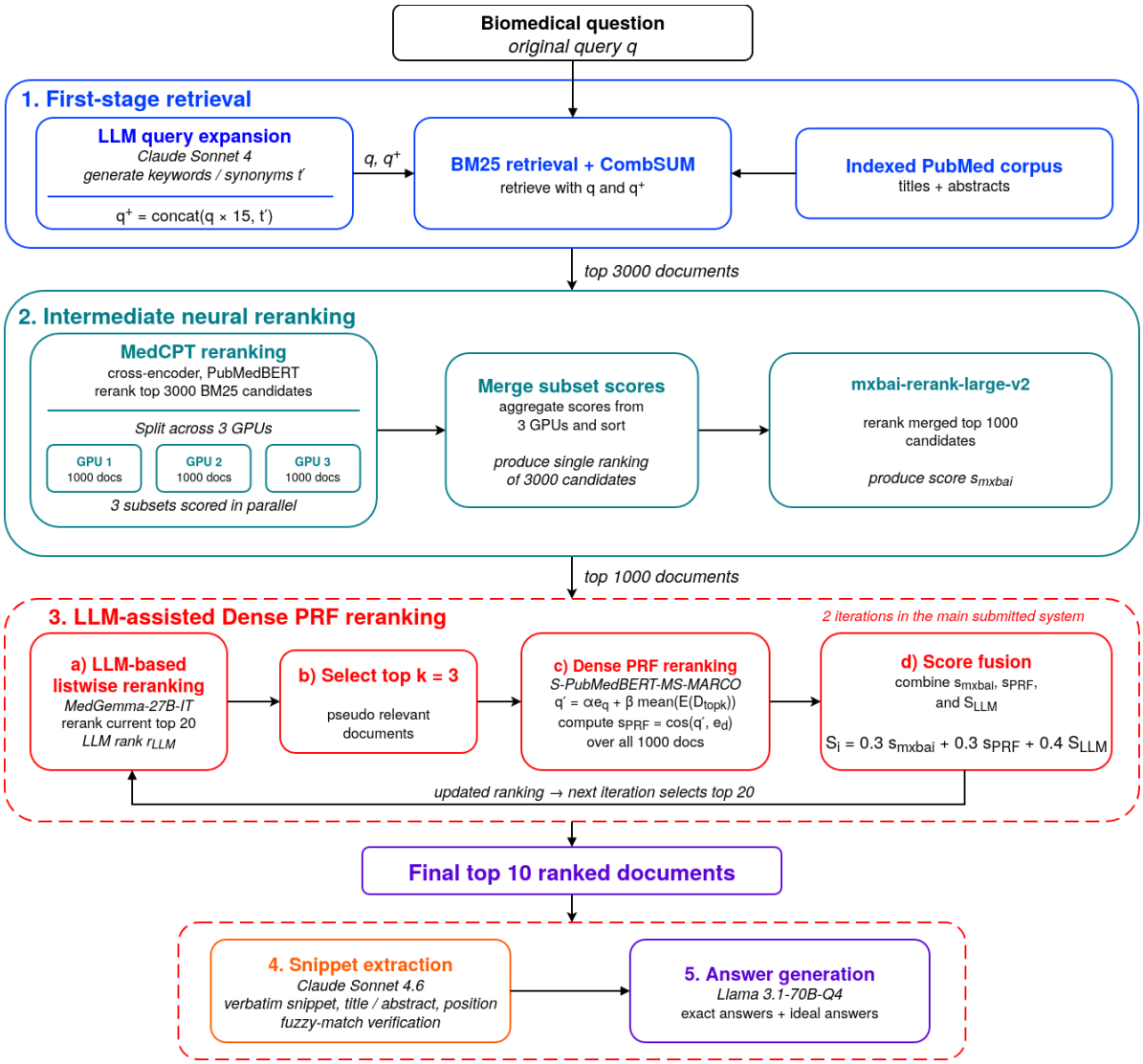

Figure 1: Overview of the MedCascade architecture.

3.2. Intermediate Neural Reranking

After first-stage retrieval, MedCascade applies two neural reranking stages to improve the precision of the candidate set while preserving sufficient recall for the later stages. We first rerank the top 3000 BM25 candidates using the MedCPT reranker [51]. MedCPT is a cross-encoder initialized from PubMedBERT and trained for biomedical query-document matching. Because scoring 3000 documents with a cross-encoder is computationally demanding, we parallelize this step across three GPUs by splitting the candidate set into three subsets of 1000 documents. The use of 1000-document candidate sets is consistent with prior studies on multi-stage retrieval [13, 52]. Each subset is scored independently, and the resulting scores are merged to produce a single ranking.

We retain the top 1000 documents after MedCPT reranking and apply a second reranking stage using mxbai-rerank-large-v2, hereafter referred to as mxbai [53, 54]. For each candidate document d_i, this stage produces a reranking score s_i^{mxbai}, which is later used in the final score fusion step. This model is a 1.5B-parameter reranker based on a Qwen2.5 architecture and has shown strong performance on English retrieval benchmarks [54]. The resulting top 1000 documents are used as a refined candidate pool for the final LLM-assisted Dense PRF reranking stage. This design keeps a relatively large set of potentially relevant documents while improving the precision of the candidates passed to the more expensive final reranking module.

3.3. LLM-assisted Dense PRF Reranking

The final retrieval stage aims to improve precision in the top-10 documents used for snippet extraction and answer generation. The module receives the top 1000 documents from the intermediate neural reranking stage. Within this candidate pool, MedCascade applies LLM-based listwise reranking only to the current top 20 documents. The highest-ranked $k = 3$ documents from this list are then treated as pseudo relevant feedback documents. These feedback documents are used to construct an updated dense query representation, which is used to rescore all 1000 candidate documents. The final ranking is obtained by combining three signals: the mxbai reranking score s_i^{mxbai}, the LLM rank signal derived from r_i^{LLM}, and the dense PRF similarity score s_i^{PRF}. The mxbai score is produced by the intermediate neural reranking stage described in Section 3.2, while the LLM rank signal and dense PRF score are introduced in Sections 3.3.1 and 3.3.2, respectively.

3.3.1. LLM-based Listwise Reranking

At each iteration, MedCascade applies listwise reranking to the current top 20 documents using the text-only version of MedGemma-27B-IT[5] [55]. MedGemma-27B-IT is a model based on Gemma 3, trained for medical text, with a long context length that allows listwise reranking using document titles and abstracts. The output of this stage is a reordered list of the top 20 documents, from which the highest-ranked documents are used as pseudo relevant feedback for the dense PRF reranking stage. For each document d_i included in this list, we denote its LLM-assigned rank as r_i^{LLM}. In our submitted configuration, we use the top $k = 3$ documents from the LLM ranking.

3.3.2. Dense PRF Reranking

We use S-PubMedBERT-MS-MARCO [56] as the bi-encoder for the dense PRF reranking stage. The top $k = 3$ documents selected by the LLM-based listwise reranker are treated as pseudo relevant feedback documents. For each question, we fetch the pre-computed embeddings of these documents and use them to construct an updated query representation. Specifically, we compute a Rocchio-like query representation as:

$$q' = \alpha e_q + \beta \,\text{mean}(E(D_{\text{top}k}))$$

where e_q is the embedding of the original question, $E(D_{\text{top}k})$ denotes the embeddings of the selected pseudo relevant documents, $\alpha = 0.3$, and $\beta = 0.7$. The document embeddings are computed using the full titles and abstracts of the selected pseudo relevant documents. We then compute the dense PRF score for each candidate document d_i in the top-1000 set as $s_i^{\text{PRF}} = \cos(q', e_{d_i})$, where e_{d_i} is the embedding of d_i. This score is used as one of the signals in the final score fusion step.

3.3.3. Score Fusion

The final ranking is obtained by combining three relevance signals for each candidate document d_i. The first signal is the mxbai reranking score s_i^{mxbai}, produced by mxbai-rerank-large-v2. The second signal is the dense PRF score s_i^{PRF}, computed from the similarity between the updated query representation and the document embedding. The third signal is the LLM rank signal derived from the MedGemma-27B-IT listwise ranking of the top 20 documents. The LLM rank signal is nonzero only for documents included in the LLM-reranked top 20 list.

Since the mxbai and dense PRF scores are produced on different scales, we apply min-max normalization to both scores for each query. For the LLM signal, we use a reciprocal-rank score:

$$s_i^{\text{LLM}} = \begin{cases} 1/r_i^{\text{LLM}}, & \text{if } d_i \text{ was ranked by the LLM,} \\ 0, & \text{otherwise.} \end{cases}$$

[5] MedGemma-27B-IT model, accessed on 2/7/26: https://huggingface.co/google/medgemma-27b-it

The final score is computed as:

$$S_i = w_{\text{mxbai}}\hat{s}_i^{\text{mxbai}} + w_{\text{PRF}}\hat{s}_i^{\text{PRF}} + w_{\text{LLM}}s_i^{\text{LLM}}.$$

We set $w_{\text{mxbai}} = 0.3$, $w_{\text{PRF}} = 0.3$, and $w_{\text{LLM}} = 0.4$, which achieved the best performance in our preliminary experiments. Documents are ranked in descending order of S_i.

3.3.4. Iterative Refinement

The LLM-assisted dense PRF reranking module can be applied iteratively. After computing the final score S_i, the candidate documents are reordered according to this score. The top 20 documents from the updated ranking are then used as input to the next iteration of the same module. In each iteration, MedGemma-27B-IT reranks the current top 20 documents, the top $k = 3$ documents are selected as pseudo relevant feedback, and the dense PRF score is recomputed over the top-1000 candidate set. In our main submitted MedCascade configuration, we used two iterations of this module, as this setting achieved the best performance in our preliminary experiments.

3.4. Snippet Extraction

For snippet extraction, we used Claude Sonnet 4.6 to identify the most relevant verbatim snippet for each retrieved document. For each question-document pair, the model was instructed to return the exact snippet text, the source section, namely title or abstract, and the snippet position in the document. Because the submitted snippets must correspond to text spans in the original source, we applied a post-processing step to verify the model output against the document text. This step uses sliding-window fuzzy matching to locate the closest matching span and correct the extracted snippet when needed.

3.5. Answer Generation

For answer generation, we used Llama-3.1-70B-Instruct[6] to produce both exact and ideal answers. The model was prompted differently depending on the question type and was provided with either the most relevant snippets or the full abstracts as context. For exact answers, the model was instructed to return the required answer format, such as a yes/no answer for yes/no questions or a list of entities for list questions. For ideal answers, the model was instructed to generate a concise paragraph-length response grounded in the provided evidence.

4. Results

The results are organized according to the BioASQ task 14B phases, leveraging the provided datasets [57, 10]. Phase A evaluates retrieval effectiveness, requiring systems to return ranked lists of up to 10 relevant documents and snippets for each biomedical question. Phase A+ evaluates exact and ideal answer generation for the same questions, using the evidence retrieved by the submitted system. Phase B evaluates answer generation when expert-selected gold articles and snippets are provided, allowing us to assess the generation component under stronger evidence conditions.

Exact answers are evaluated according to the question type, including yes/no, factoid, and list questions. Ideal answers are paragraph-sized summaries intended to approximate expert-written biomedical answers. We therefore report document and snippet retrieval results for Phase A, and exact and ideal answer results for Phases A+ and B.

4.1. Experiments Submitted

We submitted three systems to BioASQ task 14B, namely two MedCascade variants and one simple reference system. The MedCascade variants share the same overall pipeline, but differ in the configuration

[6]Llama-3.1-70B-Instruct model, accessed on 2/7/26: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

of the LLM-assisted Dense PRF reranking stage and in the evidence provided to the generation model. The reference system was included as an initial comparison point against a lighter retrieval pipeline based on BM25 and SPLADE-v3[7]. Table 1 summarizes the submitted systems, their configuration, and the BioASQ batches and phases in which they were submitted.

Table 1
Overview of the submitted systems for BioASQ 14B.

System	Submission name	Batches	Phases	Main configuration
BM25 & SPLADE	bm25 & splade	2	A, B	BM25 is used as the first-stage retriever to obtain the top 1000 candidate documents, which are then reranked with SPLADE-v3. Snippet extraction is performed through BM25-based sentence-level passage retrieval. For generation, the system uses Llama-3.1-70B-Instruct in a zero-shot setting.
MedCascade-v1	multi-stage rank&llm	3, 4	A, A+, B	MedCascade with $k = 3$ pseudo relevant documents and $r = 2$ Dense PRF iterations. Retrieved snippets are used as context for generation.
MedCascade-v2	multi-stage rank&ll	4	A, B	MedCascade with $k = 10$ pseudo relevant documents and $r = 3$ Dense PRF iterations. Document titles and abstracts are used as context for generation.

The following sections report the performance of the submitted systems across the corresponding BioASQ task 14B batches and phases. We focus on document retrieval, snippet retrieval, exact answer generation, and ideal answer generation. Since the submitted systems were not evaluated on exactly the same batches and phases, the results should be interpreted according to the submission coverage reported in Table 1.

4.2. Phase A: Document and Snippet Retrieval

Phase A evaluates the ability of the submitted systems to retrieve relevant PubMed documents and snippets. We report document retrieval results in Table 2 and snippet retrieval results in Table 3.

For each batch, we also include the top-performing system as a reference point. In both tables, rank refers to the position of each submitted system according to MAP, which is the main official metric for Phase A document and snippet retrieval. We also report mean precision, recall, F-measure, and GMAP to provide a complementary view of retrieval effectiveness.

In terms of MAP, MedCascade-v1 ranked 9th and 4th for document retrieval in batches 3 and 4, respectively, and 8th and 7th for snippet retrieval, while MedCascade-v2 ranked 20th for document retrieval and 11th for snippet retrieval in batch 4.

Table 2
Phase A document retrieval results.

Batch	System's Ranking	System	Mean precision	Recall	F-measure	MAP	GMAP
2	1st	dictycite-max	.080	.3214	.1197	.2395	.0018
2	12th	bm25 & splade	.0575	.2464	.0881	.1947	.0007
3	1st	Fleming-5	.0783	.3206	.1178	.2114	.0026
3	9th	MedCascade-v1	.0800	.3490	.1238	.1764	.0032
4	1st	dictycite-max-rew-si	.0967	.3698	.1425	.2310	.0031
4	4th	MedCascade-v1	.0900	.3644	.1344	.2034	.0028
4	20th	MedCascade-v2	.0900	.3406	.1332	.1852	.0021

[7]SPLADE-v3 model, accessed on 2/7/26: https://huggingface.co/naver/splade-v3

Table 3
Phase A snippet retrieval results.

Batch	System's Ranking	System	Mean precision	Recall	F-measure	MAP	GMAP
2	1st	RMC_2	.0373	.0846	.0419	.2108	.0001
2	34th	bm25 & splade	.0438	.0682	.0469	.0545	.0001
3	1st	UR-IW-4	.0690	.1524	.0780	.1440	.0005
3	8th	MedCascade-v1	.0560	.1425	.0687	.0954	.0010
4	1st	UR-IW-2	.0774	.2055	.0992	.1750	.0012
4	7th	MedCascade-v1	.0814	.2441	.1062	.1344	.0020
4	11th	MedCascade-v2	.0794	.2300	.1037	.1205	.0015

The BM25 & SPLADE reference system was submitted in batch 2 as an initial system to assess the difficulty of the retrieval task. It ranked 12th by MAP for document retrieval, with a MAP of .1947 compared with .2395 for the top-ranked run. However, its recall was lower than the top-ranked run, with .2464 compared with .3214. The snippet retrieval results showed a similar limitation, as the system ranked 34th by MAP. By analyzing the preliminary results and the available gold labels for this batch, we observed that this initial pipeline often failed to retrieve enough relevant evidence. This motivated the development of MedCascade as a higher-recall multi-stage retrieval pipeline.

The later MedCascade submissions reflect this recall-oriented design. In batch 3, MedCascade-v1 ranked 9th by MAP but achieved the second-best recall among the submitted systems in the task. This supports the design goal of increasing the coverage of relevant documents before applying more selective reranking stages.

In batch 4, MedCascade-v1 achieved its best Phase A document retrieval rank among our submitted runs, ranking 4th by MAP and reaching recall close to the top-ranked run. By contrast, MedCascade-v2 did not improve over MedCascade-v1.

This suggests that increasing both the number of pseudo relevant documents and the number of Dense PRF iterations did not improve the final ranking, possibly because the additional feedback introduced noise into the refinement process.

For snippet retrieval, MedCascade uses the same snippet extraction procedure across variants. Given the retrieved and reranked documents, Claude Sonnet 4.6 is used to extract the most relevant verbatim snippet for each question-document pair, including the corresponding section and position in the source document. The extracted snippets are then verified against the original document text using sliding-window fuzzy matching. Therefore, differences between MedCascade-v1 and MedCascade-v2 in Table 3 mainly reflect differences in the document rankings produced before snippet extraction. In batch 3, MedCascade-v1 ranked 8th by MAP for snippet retrieval. In batch 4, MedCascade-v1 ranked 7th, while MedCascade-v2 ranked 11th, following the same trend observed in document retrieval. Since the MedCascade variants used the same snippet extraction procedure, these differences are mainly attributable to the quality of the retrieved and reranked document lists provided to the snippet extraction module.

Overall, the Phase A results suggest that MedCascade-v1 provided the strongest retrieval configuration among our submitted systems, while MedCascade-v2 shows that increasing feedback size and iteration depth does not necessarily improve retrieval or snippet effectiveness. Its recall-oriented design was effective in retrieving relevant evidence, as shown by the second-best recall in batch 3 and recall close to the top-ranked run in batch 4.

4.3. Phase A+: Answer Generation with Retrieved Evidence

Phase A+ evaluates the ability of the submitted systems to generate exact and ideal answers using the evidence retrieved by their own pipeline. This setting combines retrieval and generation, and therefore reflects the end-to-end behavior of the system. We report exact answer results in Table 4 and ideal

Table 4
Phase A+ exact answer results.

			Yes/No				Factoid			List		
Batch	System's Ranking	System	Accuracy	F1 Yes	F1 No	Macro F1	Strict Acc.	Lenient Acc.	MRR	Mean Prec.	Recall	F-measure
3	1st	pancras_crag	1.0000	1.0000	1.0000	1.0000	.3529	.3529	.3529	.1495	.2382	.1716
3	3rd	MedCascade-v1	1.0000	1.0000	1.0000	1.0000	.3529	.4706	.4118	.2914	.3504	.3000
4	1st	ASK-Baseline	.9375	.9474	.9231	.9352	.2727	.3636	.3030	.1378	.5528	.2040
4	19th	MedCascade-v1	.8750	.9000	.8333	.8667	.0909	.2727	.1439	.2768	.5427	.3454

Table 5
Phase A+ ideal answer results using automatic evaluation scores.

Batch	System's Ranking	System	R-2 Rec.	R-2 F1	R-SU4 Rec.	R-SU4 F1
3	1st	ASK-Baseline	.2373	.0995	.2661	.1150
3	3rd	MedCascade-v1	.2372	.1191	.2445	.1267
4	1st	MedCascade-v1	.2439	.1227	.2562	.1319
4	2nd	UR-IW-5	.2377	.1144	.2614	.1285

answer results in Table 5. For exact answers, the evaluation is organized by question type, including yes/no, factoid, and list questions. For ideal answers, we report automatic ROUGE-based scores, namely ROUGE-2 and ROUGE-SU4 recall and F1.

In Phase A+, MedCascade-v1 ranked 3rd in batch 3 for exact answers and also ranked 3rd for ideal answers. For exact answers in batch 3, MedCascade-v1 matched the top-ranked system on yes/no questions and obtained higher factoid lenient accuracy, factoid MRR, and list-answer scores than the top-ranked system reported in the table. In batch 4, MedCascade-v1 ranked 19th for exact answers, but ranked 1st for ideal answers according to the automatic evaluation scores. This suggests that the retrieved evidence used by MedCascade-v1 was particularly useful for paragraph-level answer generation. At the same time, exact answer generation remained more sensitive to question type and output format, especially for factoid and list questions.

Overall, the Phase A+ results show that MedCascade-v1 can support strong end-to-end ideal answer generation when using its own retrieved evidence as context. However, exact answer generation remains more variable, suggesting that future improvements should focus not only on evidence retrieval, but also on answer type recognition and constrained answer formatting.

4.4. Phase B: Answer Generation with Expert-selected Evidence

Phase B evaluates answer generation when systems are provided with expert-selected gold articles and snippets. Unlike Phase A+, this setting reduces the influence of the system's own retrieval pipeline and allows us to examine the generation component under stronger evidence conditions. We report exact answer results in Table 6 and ideal answer results in Table 7. As in Phase A+, exact answers are evaluated by question type, while ideal answers are evaluated using automatic ROUGE-based scores.

Table 6
Phase B exact answer results.

			Yes/No				Factoid			List		
Batch	System's Ranking	System	Accuracy	F1 Yes	F1 No	Macro F1	Strict Acc.	Lenient Acc.	MRR	Mean Prec.	Recall	F-measure
2	1st	dictycite-baseline	1.0000	1.0000	1.0000	1.0000	.3000	.3000	.3000	.4465	.4280	.4301
2	28th	bm25 & splade	.9048	.9231	.8750	.8990	.3000	.3000	.3000	.4309	.4405	.4268
3	1st	MedQA-1	1.0000	1.0000	1.0000	1.0000	.5294	.5294	.5294	.4475	.4336	.4275
3	37th	MedCascade-v1	.9091	.9333	.8571	.8952	.4706	.5294	.5000	.4644	.5666	.4986
4	1st	IR_Y-1	1.0000	1.0000	1.0000	1.0000	.0909	.1818	.1212	.3443	.4239	.3661
4	10th	MedCascade-v1	.9375	.9474	.9231	.9352	.2727	.3636	.2955	.4532	.6747	.5251
4	13th	MedCascade-v2	.9375	.9474	.9231	.9352	.3636	.4545	.4091	.3912	.6380	.4609

In Phase B, the BM25 & SPLADE reference system ranked 28th for exact answers and 31st for

Table 7
Phase B ideal answer results using automatic evaluation scores.

Batch	System's Ranking	System	R-2 Rec.	R-2 F1	R-SU4 Rec.	R-SU4 F1
2	1st	Organization name	.3923	.2082	.3986	.2103
2	31st	bm25 & splade	.2663	.2176	.2702	.2156
3	1st	Organization name	.4028	.2397	.4155	.2458
3	3rd	MedCascade-v1	.3381	.1897	.3320	.1856
4	1st	Fleming-1	.3743	.1803	.3798	.1805
4	8th	MedCascade-v2	.3149	.1527	.3108	.1525
4	9th	MedCascade-v1	.3100	.1711	.3004	.1686

ideal answers in batch 2. MedCascade-v1 ranked 37th for exact answers in batch 3, but ranked 3rd for ideal answers. This suggests that the generation pipeline was more effective for paragraph-level answer generation than for exact answer formatting in this batch. In batch 4, MedCascade-v1 and MedCascade-v2 achieved the same yes/no performance. MedCascade-v2 obtained higher factoid scores, while MedCascade-v1 obtained higher list-answer scores. For ideal answers, MedCascade-v2 ranked 8th and MedCascade-v1 ranked 9th, with MedCascade-v1 obtaining a higher ROUGE-2 F1 score and MedCascade-v2 obtaining a higher ROUGE-SU4 recall score.

Overall, the Phase B results suggest that the generation component can produce competitive ideal answers when expert-selected evidence is available. However, exact answer performance remains more variable across question types, indicating that future improvements should focus on answer type control and constrained answer formatting.

5. Conclusions and Future Work

In this work, we presented MedCascade, a multi-stage biomedical question answering system developed for BioASQ task 14B. The system combines LLM-based query expansion, sparse retrieval, neural reranking, LLM-based listwise reranking, Dense PRF reranking, snippet extraction, and LLM-based answer generation. Our results show that MedCascade-v1 was the strongest configuration among our submitted systems, especially in Phase A, where it achieved strong recall and competitive MAP rankings for document and snippet retrieval. The results also show that increasing the number of pseudo relevant documents and Dense PRF iterations in MedCascade-v2 did not improve performance, suggesting that feedback quality may be more important than feedback quantity in this setting. For answer generation, MedCascade showed stronger performance for ideal answers than for exact answers, indicating that the retrieved evidence was useful for paragraph-level biomedical answer generation, while exact answer generation remains more sensitive to question type and output constraints.

Future work will focus on improving the control of the final reranking and generation stages. First, we plan to study more robust strategies for selecting pseudo relevant documents, in order to reduce the risk of introducing noisy feedback during Dense PRF reranking. Second, we aim to improve exact answer generation through question-type-specific prompting, answer validation, and constrained decoding or post-processing. Finally, we plan to investigate more efficient variants of MedCascade that preserve retrieval effectiveness while reducing the computational cost of the intermediate neural reranking approaches.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly for paraphrasing, rewording, grammar checking, and spelling correction. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. V. Loukachevitch, I. Rozhkov, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, D. Dimitriadis, A. Bekiaridou, A. Samaras, V. Patsiou, G. M. D. Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Bioasq at CLEF2026: the fourteenth edition of the large-scale biomedical semantic indexing and question answering challenge, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 - April 2, 2026, Proceedings, Part IV, Lecture Notes in Computer Science, Springer, 2026, pp. 315–324. URL: https://doi.org/10.1007/978-3-032-21321-1_42. doi:10.1007/978-3-032-21321-1_42.

[2] 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.

[3] S. Tian, Q. Jin, L. Yeganova, P. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. C. Comeau, R. Islamaj, A. Kapoor, X. Gao, Z. Lu, Opportunities and challenges for chatgpt and large language models in biomedicine and health, Briefings Bioinform. 25 (2024). URL: https://doi.org/10.1093/bib/bbad493. doi:10.1093/BIB/BBAD493.

[4] Y. Wang, Y. Zhao, L. R. Petzold, Are large language models ready for healthcare? A comparative study on clinical language understanding, in: K. Deshpande, M. Fiterau, S. Joshi, Z. C. Lipton, R. Ranganath, I. Urteaga, S. Yeung (Eds.), Machine Learning for Healthcare Conference, MLHC 2023, 11-12 August 2023, New York, USA, Proceedings of Machine Learning Research, PMLR, 2023, pp. 804–823. URL: https://proceedings.mlr.press/v219/wang23c.html.

[5] R. Pradeep, S. Sharifymoghaddam, J. Lin, Rankvicuna: Zero-shot listwise document reranking with open-source large language models, CoRR abs/2309.15088 (2023). URL: https://doi.org/10.48550/arXiv.2309.15088. doi:10.48550/ARXIV.2309.15088. arXiv:2309.15088.

[6] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren, Is chatgpt good at search? investigating large language models as re-ranking agents, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 14918–14937. URL: https://doi.org/10.18653/v1/2023.emnlp-main.923. doi:10.18653/V1/2023.EMNLP-MAIN.923.

[7] Y. Gao, L. Zong, Y. Li, Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 117–129. URL: https://ceur-ws.org/Vol-3740/paper-10.pdf.

[8] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of bioasq tasks 13b and synergy13 in clef2025, in: CLEF, 2025.

[9] W. Xie, Y. Fang, X. Zheng, D. Liu, Z. Li, When retrieval hurts: A critical analysis of rag in medical question answering, in: Proceedings of the 2025 3rd International Conference on Artificial Intelligence, Systems and Network Security, 2025, pp. 312–317.

[10] 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 2025 Working Notes, 2026.

[11] M. Luo, A. Mitra, T. Gokhale, C. Baral, Improving biomedical information retrieval with neural re-

trievers, CoRR abs/2201.07745 (2022). URL: https://arxiv.org/abs/2201.07745. arXiv:2201.07745.
[12] A. Rosso-Mateus, F. A. González, M. Montes, Mindlab neural network approach at bioasq 6b, in: Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, 2018, pp. 40–46.
[13] A. Shin, Q. Jin, Z. Lu, Multi-stage literature retrieval system trained by pubmed search logs for biomedical question answering, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023, pp. 178–189. URL: https://ceur-ws.org/Vol-3497/paper-016.pdf.
[14] E. Sokli, P. Kasela, G. Peikos, G. Pasi, Investigating mixture of experts in dense retrieval, 2024. URL: https://arxiv.org/abs/2412.11864. arXiv:2412.11864.
[15] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., Okapi at trec (1994).
[16] T. Formal, B. Piwowarski, S. Clinchant, SPLADE: sparse lexical and expansion model for first stage ranking, CoRR abs/2107.05720 (2021). URL: https://arxiv.org/abs/2107.05720. arXiv:2107.05720.
[17] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550/. doi:10.18653/v1/2020.emnlp-main.550.
[18] S. Verma, F. Jiang, X. Xue, Beyond retrieval: Ensembling cross-encoders and GPT rerankers with llms for biomedical QA, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 666–680. URL: https://ceur-ws.org/Vol-4038/paper_50.pdf.
[19] V. Novotný, M. Stefánik, Combining sparse and dense information retrieval, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, 2022, pp. 104–118. URL: https://ceur-ws.org/Vol-3180/paper-06.pdf.
[20] M. Weber, E. Davis, R. Johnson, S. Wilson, Joint optimization framework combining sparse representation and dense retrieval in pubmed retrieval, International Journal of Computational and Biological Sciences 3 (2026) 77–86.
[21] S. G. Akkuş, H. Emekci, Ranking matters: A comparative study of bm25, bert-e5, and hybrid (bm25 + bert-e5) retrival systems on the squad 2.0 dataset, in: 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), 2025, pp. 1–6. doi:10.1109/ACDSA65407.2025.11166004.
[22] G. Rosa, L. H. Bonifacio, V. Jeronymo, H. Q. Abonizio, M. Fadaee, R. A. Lotufo, R. Nogueira, In defense of cross-encoders for zero-shot retrieval, CoRR abs/2212.06121 (2022). URL: https://doi.org/10.48550/arXiv.2212.06121. doi:10.48550/ARXIV.2212.06121. arXiv:2212.06121.
[23] T. M. Almeida, R. A. A. Jonker, J. A. Reis, J. R. Almeida, S. Matos, BIT.UA at bioasq 12: From retrieval to answer generation, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 47–67. URL: https://ceur-ws.org/Vol-3740/paper-05.pdf.
[24] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/V1/D19-1410.
[25] D. Chen, S. Zhang, X. Zhang, K. Yang, Cross-lingual passage re-ranking with alignment augmented multilingual BERT, IEEE Access 8 (2020) 213232–213243. URL: https://doi.org/10.1109/ACCESS.

2020.3041605. doi:10.1109/ACCESS.2020.3041605.

[26] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. J. Orr, L. Zheng, M. Yüksekgönül, M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, CoRR abs/2211.09110 (2022). URL: https://doi.org/10.48550/arXiv.2211.09110. doi:10.48550/ARXIV.2211.09110. arXiv:2211.09110.

[27] S. Zhuang, B. Liu, B. Koopman, G. Zuccon, Open-source large language models are strong zero-shot query likelihood models for document ranking, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, Findings of ACL, Association for Computational Linguistics, 2023, pp. 8807–8817. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.590. doi:10.18653/V1/2023.FINDINGS-EMNLP.590.

[28] H. Yang, S. Li, T. Gonçalves, Enhancing biomedical question answering with large language models, Inf. 15 (2024) 494. URL: https://doi.org/10.3390/info15080494. doi:10.3390/INFO15080494.

[29] X. Ma, X. Zhang, R. Pradeep, J. Lin, Zero-shot listwise document reranking with a large language model, CoRR abs/2305.02156 (2023). URL: https://doi.org/10.48550/arXiv.2305.02156. doi:10.48550/ARXIV.2305.02156. arXiv:2305.02156.

[30] S. Zhuang, H. Zhuang, B. Koopman, G. Zuccon, A setwise approach for effective and highly efficient zero-shot ranking with large language models, in: G. H. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, Y. Zhang (Eds.), Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, ACM, 2024, pp. 38–47. URL: https://doi.org/10.1145/3626772.3657813. doi:10.1145/3626772.3657813.

[31] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) 267–270. URL: https://doi.org/10.1093/nar/gkh061. doi:10.1093/NAR/GKH061.

[32] G. Peikos, D. Alexander, G. Pasi, A. P. de Vries, Investigating the impact of query representation on medical information retrieval, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II, Lecture Notes in Computer Science, Springer, 2023, pp. 512–521. URL: https://doi.org/10.1007/978-3-031-28238-6_42. doi:10.1007/978-3-031-28238-6_42.

[33] O. Vechtomova, Y. Wang, A study of the effect of term proximity on query expansion, J. Inf. Sci. 32 (2006) 324–333. URL: https://doi.org/10.1177/0165551506065787. doi:10.1177/0165551506065787.

[34] G. Peikos, S. Symeonidis, P. Kasela, G. Pasi, Utilizing chatgpt to enhance clinical trial enrollment, CoRR abs/2306.02077 (2023). URL: https://doi.org/10.48550/arXiv.2306.02077. doi:10.48550/ARXIV.2306.02077. arXiv:2306.02077.

[35] G. Peikos, P. Kasela, G. Pasi, Leveraging large language models for medical information extraction and query generation, in: IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2024, Bangkok, Thailand, December 9-12, 2024, IEEE, 2024, pp. 367–372. URL: https://doi.org/10.1109/WI-IAT62293.2024.00058. doi:10.1109/WI-IAT62293.2024.00058.

[36] X. Chen, X. Chen, B. He, T. Wen, L. Sun, Analyze, generate and refine: Query expansion with llms for zero-shot open-domain QA, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Findings of ACL, Association for Computational Linguistics, 2024, pp. 11908–11922. URL: https://doi.org/10.18653/v1/2024.findings-acl.708. doi:10.18653/V1/2024.FINDINGS-ACL.708.

[37] F. Borazio, A. Shcherbakov, D. Croce, R. Basili, Unitor at bioasq 2025: Modular biomedical QA with synthetic snippets and multiple task answer generation, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina

(Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 165–197. URL: https://ceur-ws.org/Vol-4038/paper_10.pdf.

[38] Z. A. Nazi, V. Hristidis, A. L. McLean, J. A. Meem, M. T. A. Chowdhury, Ontology-guided query expansion for biomedical document retrieval using large language models, CoRR abs/2508.11784 (2025). URL: https://doi.org/10.48550/arXiv.2508.11784. doi:10.48550/ARXIV.2508.11784. arXiv:2508.11784.

[39] J. Xu, W. B. Croft, Query expansion using local and global document analysis, in: H. Frei, D. Harman, P. Schäuble, R. Wilkinson (Eds.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum), ACM, 1996, pp. 4–11. URL: https://doi.org/10.1145/243199.243202. doi:10.1145/243199.243202.

[40] A. Arampatzis, G. Peikos, S. Symeonidis, Pseudo relevance feedback optimization, Inf. Retr. J. 24 (2021) 269–297. URL: https://doi.org/10.1007/s10791-021-09393-5. doi:10.1007/S10791-021-09393-5.

[41] T. Sakai, T. Manabe, M. Koyama, Flexible pseudo-relevance feedback via selective sampling, ACM Trans. Asian Lang. Inf. Process. 4 (2005) 111–135. URL: https://doi.org/10.1145/1105696.1105699. doi:10.1145/1105696.1105699.

[42] X. Wang, C. MacDonald, N. Tonellotto, I. Ounis, Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval, ACM Trans. Web 17 (2023) 3:1–3:39. URL: https://doi.org/10.1145/3572405. doi:10.1145/3572405.

[43] H. Li, S. Zhuang, A. Mourad, X. Ma, J. Lin, G. Zuccon, Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I, Lecture Notes in Computer Science, Springer, 2022, pp. 599–612. URL: https://doi.org/10.1007/978-3-030-99736-6_40. doi:10.1007/978-3-030-99736-6_40.

[44] H. Li, A. Mourad, S. Zhuang, B. Koopman, G. Zuccon, Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls, ACM Trans. Inf. Syst. 41 (2023) 62:1–62:40. URL: https://doi.org/10.1145/3570724. doi:10.1145/3570724.

[45] H. Kim, H. Lee, Y. Cho, J. Park, J. Park, S. Park, Y. T. Chok, S. Baek, D. Lee, J. Kang, Prompting matters: Snippet-aware strategies for biomedical QA with llms in bioasq 13b, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 373–382. URL: https://ceur-ws.org/Vol-4038/paper_25.pdf.

[46] D. Galat, D. M. Aliod, LLM ensemble for RAG: role of context length in zero-shot question answering for bioasq challenge, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 266–272. URL: https://ceur-ws.org/Vol-4038/paper_16.pdf.

[47] W. Zhou, T. H. Ngo, Using pretrained large language model with prompt engineering to answer biomedical questions, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 253–268. URL: https://ceur-ws.org/Vol-3740/paper-22.pdf.

[48] L. Wang, N. Yang, F. Wei, Query2doc: Query expansion with large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 9414–9423. URL: https://doi.org/10.18653/v1/2023.emnlp-main.585. doi:10.18653/V1/2023.EMNLP-MAIN.585.

[49] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, M. Bendersky, Query expansion by prompting large language models, CoRR abs/2305.03653 (2023). URL: https://doi.org/10.48550/arXiv.2305.03653.

doi:10.48550/ARXIV.2305.03653. arXiv:2305.03653.
[50] E. A. Fox, J. A. Shaw, Combination of multiple searches, NIST special publication SP 243 (1994).
[51] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, Bioinform. 39 (2023). URL: https://doi.org/10.1093/bioinformatics/btad651. doi:10.1093/BIOINFORMATICS/BTAD651.
[52] S. Chatterjee, REGENT: relevance-guided attention for entity-aware multi-vector neural re-ranking, in: Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2025, Xi'an, China, December 7-10, 2025, ACM, 2025, pp. 199–210. URL: https://doi.org/10.1145/3767695.3769476. doi:10.1145/3767695.3769476.
[53] X. Li, A. Shakir, R. Huang, J. Lipp, J. Li, Prorank: Prompt warmup via reinforcement learning for small language models reranking, CoRR abs/2506.03487 (2025). URL: https://doi.org/10.48550/arXiv.2506.03487. doi:10.48550/ARXIV.2506.03487. arXiv:2506.03487.
[54] S. Lee, R. Huang, A. Shakir, J. Lipp, Baked-in brilliance: Reranking meets rl with mxbai-rerank-v2, https://www.mixedbread.com/blog/mxbai-rerank-v2, 2025.
[55] G. Research, G. DeepMind, Medgemma technical report, CoRR abs/2507.05201 (2025). URL: https://doi.org/10.48550/arXiv.2507.05201. doi:10.48550/ARXIV.2507.05201. arXiv:2507.05201.
[56] P. Deka, A. Jurek-Loughrey, P. Deepak, Improved methods to aid unsupervised evidence-based fact checking for online health news, Journal of Data Intelligence 3 (2022) 474–504.
[57] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.