LLM Ensemble for RAG: Role of Context Length in Zero-Shot Question Answering for BioASQ Challenge

Notebook for the BioASQ Lab at CLEF 2025

Dima Galat[1], Diego Molla-Aliod[2]

[1]University of Technology Sydney (UTS), Australia
[2]Macquarie University, Australia

Abstract

Biomedical question answering (QA) poses significant challenges due to the need for precise interpretation of specialized knowledge drawn from a vast, complex, and rapidly evolving corpus. In this work, we explore how large language models (LLMs) can be used for information retrieval (IR), and an ensemble of zero-shot models can accomplish state-of-the-art performance on a domain-specific Yes/No QA task. Evaluating our approach on the BioASQ challenge tasks, we show that ensembles can outperform individual LLMs and in some cases rival or surpass domain-tuned systems - all while preserving generalizability and avoiding the need for costly fine-tuning or labeled data. Our method aggregates outputs from multiple LLM variants, including models from Anthropic and Google, to synthesize more accurate and robust answers. Moreover, our investigation highlights a relationship between context length and performance: while expanded contexts are meant to provide valuable evidence, they simultaneously risk information dilution and model disorientation. These findings emphasize IR as a critical foundation in Retrieval-Augmented Generation (RAG) approaches for biomedical QA systems. Precise, focused retrieval remains essential for ensuring LLMs operate within relevant information boundaries when generating answers from retrieved documents. Our results establish that ensemble-based zero-shot approaches, when paired with effective RAG pipelines, constitute a practical and scalable alternative to domain-tuned systems for biomedical question answering.

Keywords

large language model, biomedical question answering, information retrieval, natural language processing, BioASQ, CEUR-WS

1. Introduction

Biomedical Question Answering (QA) is a challenging task, and BioASQ is an annual competition aimed at fostering the development of intelligent systems specialized in Information Retrieval (IR) and QA within the biomedical domain [1]. The competition consists of three distinct phases: phase A, which focuses on a biomedical IR task; phase B, centered on QA and summarization tasks; and phase A+, which seeks to develop an end-to-end approach for both phases of the challenge combined, has been introduced for the first time in 2024 to encourage further development in this area [2]. Notably, a third of the participating teams attempted all three phases of the challenge that year.

This paper presents the following key advancements:

- We develop a zero-shot QA ensembling framework that leverages Large Language Models (LLMs) and answer synthesis to accomplish state-of-the-art results on a Yes/No QA task.
- We design a multi-stage biomedical IR pipeline that combines LLM-generated queries, BM25 lexical search, and semantic reranking to enable high-recall, high-precision document selection for Retrieval-Augmented Generation (RAG) QA in Phase A+.
- We demonstrate how context influences the QA results and show that IR is still an important part of a modern end-to-end QA solution.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
dima.galat[@]student.uts.edu.au (D. Galat); diego.molla-aliod[@]mq.edu.au (D. Molla-Aliod)
0000-0003-3825-2142 (D. Galat); 0000-0003-4973-0963 (D. Molla-Aliod)

2. Related work

IR and QA are two well-established fields in Natural Language Processing (NLP) and both of them have had extensive research for decades. Most relevant for us is work on the use of LLMs for query generation in IR, answer re-ranking after a first pass by a traditional retrieval system, and the use of LLMs and RAG for QA.

LLMs have been used to generate search queries, or to rewrite initial search queries to improve their performance. Techniques include the use of zero-shot, few-shot and chain-of-thought prompting [3, 4, 5], supervised fine-tuning [6], and reinforcement learning [7].

Search results have been reranked using a wide range of techniques, the most common being applying a similarity metric such as cosine similarity to the output of a cross-encoder or a twin model such as Sentence-BERT [8]. Google released the ReFr open-source framework for re-ranking search results, which allows researchers to explore multiple features and learning methods [9].

LLMs have also been used to generate answers to questions, either in a zero-shot, or few-shot manner, or by fine-tuning on question-answering datasets. The integration of contextual information through RAG consistently improves answer quality and factual accuracy compared to LLM approaches not relying on additional relevant context information [10]. A quick look at the proceedings of BioASQ 12 at CLEF 2024 shows that, out of a total of 23 papers, 6 papers used the terms LLM or Large Language Model in their title, and 3 additional papers used the terms RAG or Retrieval-Augmented Generation.

3. Our methodology

3.1. Information Retrieval Pipeline

To support high-quality RAG for Phase A+, we developed an IR pipeline that integrates traditional lexical search with LLM-based query generation and semantic reranking (Fig. 1).

We index all PubMed article titles and abstracts in an Elasticsearch instance, using BM25 retrieval as the ranking function. For each input question, we use Gemini 2.0 Flash to generate a structured Elasticsearch query that captures the semantic intent of the question using synonyms, related terms, and full boolean query string syntax rules supported by Elasticsearch. This query is validated using regular expressions and then is used to retrieve up to 10,000 documents.

If the initial query returns fewer than five documents, we invoke Gemini 2.5 Pro Preview (05-06) to automatically revise the query. The model is prompted to enhance retrieval recall by enabling approximate matching and omitting overly rare or domain-specific terms. This refinement step is done to improve the query coverage while maintaining relevance. Our experiments have shown that this process is required in less than 5% of the queries in the BioASQ 13 test set.

Following document retrieval, we apply a semantic reranking model (Google semantic-ranker-default-004) to reduce the number of candidate documents [11]. This model re-scores the initially retrieved documents based on semantic similarity to the original question, allowing us to select the top 300 most relevant documents. This reranked subset is used for downstream RAG-based QA, since despite really long context supported by modern Transformer architectures [12, 13], we could not get adequate QA results on full article abstracts without this step.

This multi-stage retrieval approach, combining LLM-generated queries, a traditional BM25 search, and semantic reranking, enables flexible, high-recall, and high-precision document selection tailored to complex biomedical queries.

Finally, we have added additional IR searches to handle the cases where a QA step does not return a response based on the evidence retrieved from Elasticsearch. We have observed that Elasticsearch context might not provide sufficient evidence for QA in 3-7% of test cases for Phase A+, depending on the batch. An automated process is used to expand IR sources to address these cases. First, we are using a Google search restricted to PubMed sources to attempt to find new matches. If that fails, we extend our sources to include Home of the Office of Health Promotion and Disease Prevention, WebMD,

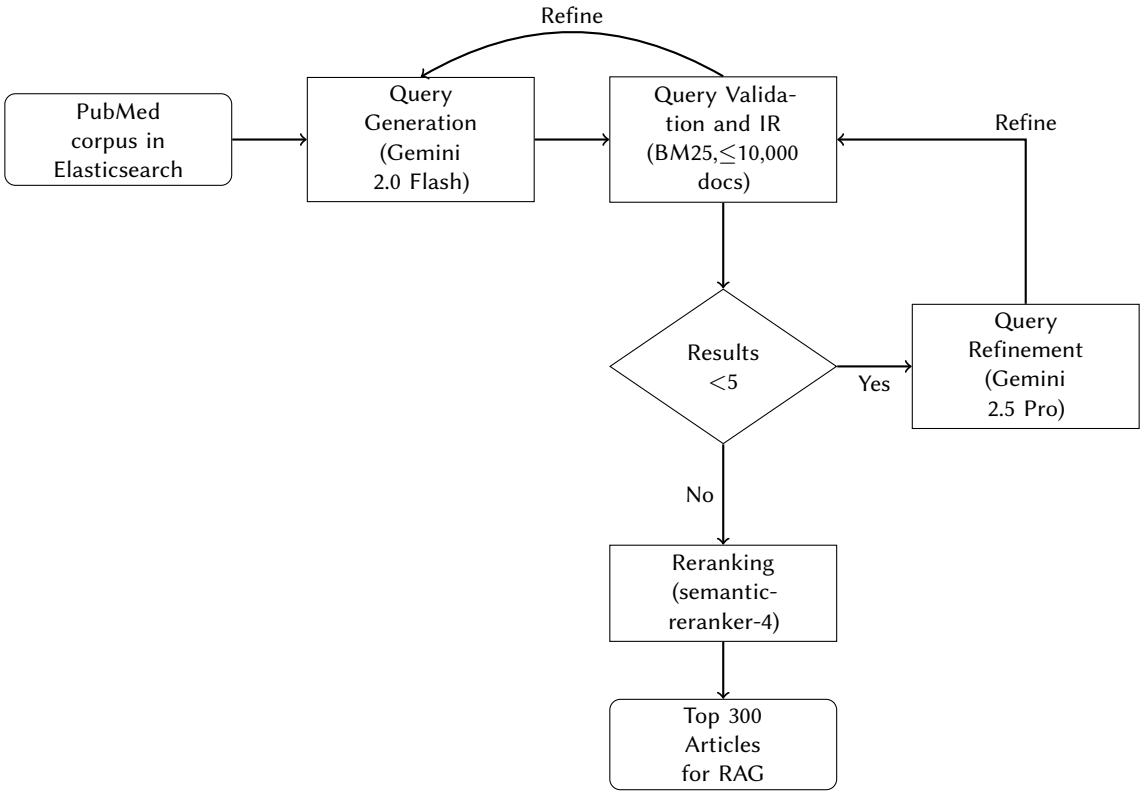

Figure 1: IR process

Healthline, and Wikipedia. This ensures that we have an answer candidate for all questions in Phase A+ test sets.

3.2. Question Answering Pipeline

We adopt a unified, zero-shot QA framework for both Phase A+ and Phase B of the challenge. While the core QA procedure remains consistent across phases, Phase A+ incorporates an additional IR step to verify the presence of candidate answers within relevant documents (described at the end of Section 3.1). This ensures that selected documents contain sufficient information to support answer generation.

The system uses zero-shot prompting, tailored to the question type: Yes/No, Factoid, or List. We experiment with multiple types of input context: (1) IR-derived results from Phase A+, (2) curated snippets provided in Phase B, and (3) full abstracts of articles selected during Phase B. This allows us to examine the influence of context granularity on answer accuracy and completeness.

To generate candidate answers, we leverage several large language models (LLMs): Gemini 2.0 Flash, Gemini 2.5 Flash Preview (2025-04-17), and Claude 3.7 Sonnet (2025-02-19). Prompts are adjusted using examples derived from the BioASQ 11 test set, improving the response structure and quality.

To consolidate candidate answers, we perform a secondary synthesis step using Gemini 2.0 Flash. This model is prompted to resolve any contradictions, select the most precise and specific answer components, and integrate complementary information into a single, unified response. As part of this step, the model also returns a confidence score estimating the reliability of the synthesized answer. If the score is below a predefined threshold (0.5, determined empirically), the synthesis is re-run with reduced sampling temperature (from 0.1 to 0.0) to improve determinism. This synthesis process is evaluated using the BioASQ 12 dataset to ensure consistency with benchmark standards.

Table 1
Results of our runs on BioASQ 13 Phase A+, Yes/No questions.

Batch	System	Accuracy	Ranking
3	Extractive	0.73	41
	(last)	0.23	58
4	Extractive	0.92	1
	Simple truncation	0.88	11
	Kmeans	0.65	67
	(last)	0.65	67

Table 2
Results of our runs on BioASQ 13 Phase A+, Factoid questions.

Batch	System	MRR	Ranking
3	Extractive	0.14	41
	(last)	0.05	47
4	Extractive	0.43	17
	Simple truncation	0.29	51
	Kmeans	0.05	62
	(last)	0.05	62

4. Results

4.1. Phase A+

BioASQ competition results are evaluated using accuracy, which measures the proportion of correctly answered yes/no questions, and MRR (Mean Reciprocal Rank), which assesses how early correct answers appear in the ranked list of factoid question responses.

Our IR pipeline has been actively developed during the BioASQ 13 competition, and as a result, it was not ready in time to submit the first two batches. Notably, on batch 4, our system achieved state-of-the-art results on Yes/No questions, underscoring the effectiveness of the RAG approach described in Section 3.1 (Table 1). Despite this success, our system encountered challenges in producing the structured output required for List and Factoid questions, which is a consistent issue we've seen with zero-shot generation. Table 2 shows the results for factoid questions. It's important note that Simple truncation is a system using Gemini 2.0 Flash described in Figure 1, and Extractive is the system using Claude 3.7 Sonnet —which worked better for long context extraction despite technically having a smaller context size available.

4.2. Phase B

We evaluated our system on the BioASQ 12 dataset and observed competitive performance across all batches. Overall, our system ranked 5th for exact answers across our experiments compared to systems presented last year, with batch 4 posing the greatest challenge across all question types. Among the models tested, Gemini consistently outperformed Claude, with Gemini 2.0 Flash showing significantly better results compared to Gemini 2.5 Flash or Pro. The system showed its strongest performance on Yes/No questions, while List-type questions were the most challenging. We found that using longer contexts —like full abstracts —generally hurts the answer quality. This can be attributed to difficulties in generating well-structured outputs for List and Factoid questions, further highlighting that answer generation and formatting should be decoupled into separate stages to improve the results on exact questions requiring more nuanced responses than a binary answer.

Across the BioASQ 13 Phase B batches, our system demonstrated consistently strong performance on Yes/No questions. In batches 1, 2, and 3 we achieved state-of-the-art results on Yes/No questions using

Table 3
Results of our runs on BioASQ 13 Phase B, Yes/No questions.

Batch	System	Accuracy	Ranking
1	Simple truncation	1.00	1
	Extractive	0.94	23
	Kmeans	0.94	23
	(last)	0.29	72
2	Simple truncation	1.00	1
	Extractive	0.88	47
	Kmeans	0.88	47
	(last)	0.47	72
3	Simple truncation	0.95	1
	Extractive	0.95	1
	Kmeans	0.95	1
	(last)	0.27	66
4	Simple truncation	0.96	3
	Extractive	0.96	3
	Kmeans	0.96	3
	(last)	0.27	79

Table 4
Results of our runs on BioASQ 13 Phase B, Factoid questions.

Batch	System	MRR	Ranking
1	Simple truncation	0.44	38
	Extractive	0.38	53
	Kmeans	0.37	54
	(last)	0.11	67
2	Simple truncation	0.57	20
	KMeans	0.54	32
	Extractive	0.52	36
	(last)	0.07	66
3	Simple truncation	0.60	2
	Kmeans	0.49	4
	Extractive	0.10	55
	(last)	0.03	59
4	Kmeans	0.50	30
	Extractive	0.48	38
	Simple truncation	0.45	46
	(last)	0.05	73

snippets provided for this stage and an ensemble of Gemini 2.0 Flash and Gemini 2.5 Flash Preview (2025-04-17) (Table 3). In batch 3, we managed to refine our prompts to also achieve a second place in Factoid questions using the same model and context selection (Table 4). In batch 4, this approach ranked third on Yes/No questions (Table 3), demonstrating the robustness of our synthesis pipeline across varied input selection strategies.

5. Conclusion and Future Work

5.1. Conclusions

This work demonstrates the effectiveness of integrating zero-shot LLMs with traditional IR systems for providing an end-to-end approach for QA. LLMs can help bridge the semantic gap between queries and relevant documents and are effective when used in a specialized domain [14]. Our findings align with the growing trend, as evidenced by the substantial adoption of LLM and RAG techniques in recent BioASQ competitions [2], as well as with studies that show that query rewriting provides substantial performance advantages [3, 5, 4].

The multi-stage IR pipeline approach provides robust performance while maintaining computational efficiency. We have shown that a combination of LLM query generation, cross-encoder re-ranking, and RAG are capable of processing very long domain-specific contexts, achieving a SOTA Yes/No QA system performance on multiple BioASQ batches this year. We found that producing structured data outputs for other question types can be challenging, especially as the context size increases.

5.2. Future Work

Several promising directions emerge from this research that warrant further investigation:

Evaluation and Robustness: Developing more comprehensive evaluation frameworks that assess not only accuracy but also consistency, bias, and hallucination rates across diverse query types and domains [15]. ARES proposes an approach for evaluating RAG systems that fine-tunes lightweight LLM judges using synthetic data and minimal human annotations, achieving high accuracy across tasks and domains while outperforming prior methods [16].

Interactive Systems: Development of conversational interfaces that can clarify ambiguous queries, and provide explanations for retrieved information and generated answers [17]. Uncertainty detection methods can be used to dynamically trigger retrieval in RAG systems, reducing unnecessary retrievals while maintaining or even improving answer quality in long-form question answering tasks [18].

6. Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly to assist with grammar and spelling checks, paraphrasing, and rewording. The authors confirm that the intellectual content, analysis, interpretations, and conclusions presented in this paper are entirely their own and take full responsibility for the publication's content.

References

[1] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artiéres, A. C. N. Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (2015). doi:10.1186/s12859-015-0564-6.

[2] A. Nentidis, A. Krithara, G. Paliouras, M. Krallinger, L. G. Sanchez, S. Lima, E. Farre, N. Loukachevitch, V. Davydova, E. Tutubalina, BioASQ at CLEF2024: The Twelfth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, 2024, pp. 490–497. doi:10.1007/978-3-031-56069-9_67.

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

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

[5] M. Alaofi, L. Gallagher, M. Sanderson, F. Scholer, P. Thomas, Can generative LLMs create query variants for test collections? an exploratory study, in: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, 2023, pp. 1869–1873.

[6] W. Peng, G. Li, Y. Jiang, Z. Wang, D. Ou, X. Zeng, D. Xu, T. Xu, E. Chen, Large language model based long-tail query rewriting in taobao search, in: Companion Proceedings of the ACM Web Conference 2024, 2024, pp. 20–28.

[7] X. Ma, Y. Gong, P. He, H. Zhao, N. Duan, Query rewriting in retrieval-augmented large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5303–5315.

[8] 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[9] D. M. Bikel, K. B. Hall, ReFr: An open-source reranker framework, in: Interspeech 2013, 2013, pp. 756–758.

[10] 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, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 9459–9474.

[11] Google Cloud, Ranking and re-ranking search results, https://cloud.google.com/generative-ai-app-builder/docs/ranking, 2024. Accessed: 2025.

[12] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, in: Advances in Neural Information Processing Systems, volume 2020-December, 2020.

[13] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer (2020).

[14] 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: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 6769–6781.

[15] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, FActScore: Fine-grained atomic evaluation of factual precision in long form text generation, in: EMNLP, 2023. URL: https://arxiv.org/abs/2305.14251.

[16] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, ARES: An automated evaluation framework for retrieval-augmented generation systems, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024, pp. 4392–4408.

[17] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, M. Iyyer, BERT with history answer embedding for conversational question answering, in: Proceedings of the 43rd International ACM SIGIR Conference, 2020, pp. 1133–1136.

[18] W. Zhang, Y. Liu, H. Chen, X. Wang, To retrieve or not to retrieve? uncertainty detection for dynamic retrieval-augmented generation, arXiv preprint arXiv:2501.09292 (2025). URL: https://arxiv.org/abs/2501.09292.