Reshaping Biomedical Scientific Literature in a RAG Pipeline for Question Answering

Maël Lesavourey[1,*], Gilles Hubert[1]

[1]IRIT, Université de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France

Abstract

Biomedical Question Answering (BQA) poses specific challenges due to the specialized vocabulary and complex semantic structures of biomedical literature. Large Language Models (LLMs) have shown great performance in several Natural Language Understanding and Generation tasks. However, their effectiveness tends to drop in domain-specific contexts such as biomedicine. Polysemy, complex lexical structures, and the need of precise and factual information exacerbate their limitations. To address these issues, Retrieval-Augmented Generation (RAG) pipelines have become a promising approach, combining the strengths of retrieval methods with LLMs to incorporate domain-specific knowledge into the generation process. In this article, we investigate the role of context in enhancing the performance of RAG pipelines for BQA. We show that incorporating a context grounded on proper literature reshaping affects positively the quality of generated answers, improving both semantic and lexical metrics. We also show that it has more effects on Precision than on Recall. This work underscores the importance of structuring appropriately the context to enhance the performance of LLMs and assist them in processing and selecting relevant information.

Keywords

Retrieval-Augmented Generation, Biomedical Question Answering, Information Retrieval, Answer Generation, Scientific Literature Processing

1. Introduction

Since their release, language models (LMs) such as BERT [1] and GPT [2] have gradually been adopted for a wide range of tasks related to Natural Language Understanding (NLU) and Processing (NLP). Their ability to understand the semantic relation between words in a document has reshaped traditional approaches in various fields like Information Retrieval (IR), achieving State-Of-The-Art (SOTA) results in multiple tasks, e.g., document ranking, classification, and text generation [3, 4]. However, those models do not perform well when applied to domain-specific corpora like biomedical literature and legal documents [5, 6]. The main reasons are the particular characteristics of those texts which amplify the semantic gap between general knowledge and specialized concepts. Biomedical literature is composed of complex lexical structures like chemical formulas, proper nouns, and abbreviations. Moreover, the understanding of such literature is harder due to its polysemy, for example the expressions "Heart Attack", "Myocardial Infarction", "Cardiovascular Stroke" having the same meaning[1].

Addressing these challenges in the context of Biomedical Question Answering (BQA) tasks requires careful consideration of the domain's specific characteristics. A wide range of BQA tasks exists [7], each one having its own particularities regarding the content of the corpus, response format and targeted audience. We consider scientific literature as our source of information, while the query and its answer should target specialized readers, and be written in natural language. This task sits at the intersection of IR and language generation.

A first method to consider the specific characteristics of biomedical corpora has been to use LMs pre-trained on such texts [8, 9, 10]. However, several works [11, 12, 13] have shown that, despite

SCOLIA '25: First International Workshop on Scholarly Information Access (SCOLIA), April 10, 2025, Lucca, Italy

*Corresponding author.

mael.lesavourey@irit.fr (M. Lesavourey); gilles.hubert@irit.fr (G. Hubert)

https://www.irit.fr/~Gilles.Hubert/ (G. Hubert)

0000-0003-3494-7561 (G. Hubert)

[1]https://meshb.nlm.nih.gov/record/ui?ui=D009203

enhanced performance, such models still lack semantic understanding. With recent large LMs (LLMs), this method is not an option because it would be highly expensive to train one model from scratch. Therefore, several methods have been proposed to address the task of knowledge incorporation into LLMs. Retrieval-Augmented Generation (RAG) [14] combines text generation with relevant document retrieval mechanisms to contextualize responses. In a different way, In-Context Learning (ICL) [15] aims at aligning the generated responses with the user's expectations by providing examples directly in the model inputs. However, the effectiveness of those approaches is dependent on which context is extracted and how it is structured, e.g., examples of pairs (query, answer), plain text from scientific publications, or semantic predications.

We study in this paper how to properly incorporate domain-specific knowledge extracted from scientific publications into LLMs in order to overcome their limitations and what is the impact of adding such context for BQA.

In the remainder of this article, we first present related works on RAG and BQA. Then, we describe the method implemented to address this task, followed by a detailed presentation of the models and technologies used for its implementation and evaluation. We then analyze the results before concluding with a discussion on the implications of our approach and future research opportunities.

2. Related Works

Our work is related to different domains, i.e, IR, LMs, and BQA, as introduced in the following sections.

2.1. Information Retrieval

First approaches in IR were based on lexical matching, using statistics measuring co-occurrences of words between several texts (e.g., a document and a query). A well-known method, BM25, is based on TF-IDF scoring and takes advantage of different concepts like term frequency, rareness, and text length to compute a similarity score. Their main limitation lies in their inability to take into account the semantic meaning of the text (e.g., use of synonyms or paraphrased terms). To this end, researchers have shown interest in developing dense retrievers [16] that capture semantic relationships that go beyond exact word matching.

2.2. Language Models

The transformer architecture was introduced in [17]. It is based on the self-attention mechanism that enables to capture both local and global dependencies of a sequence of tokens. Two major families of LMs have emerged: encoder and decoder based models. BERT [1] has been the most widely studied encoder-based model since its release. Its pre-train-then-fine-tune paradigm led to significant improvements in tasks such as text classification and named entity recognition.

In the same time, decoder-based models that focus on generating new tokens by predicting the next word of a sequence have been developed. GPT-1 [2] demonstrated that training on large corpora could produce a generative model capable of handling several language comprehension and generation tasks. GPT-2 [18] marked a breakthrough by considerably increasing the number of parameters of LLMs and the size of the training corpus. Its ability to perform different tasks without fine-tuning also marked a turning point, paving the way for ICL, which makes it possible to guide the behaviour of LLMs without fine-tuning. More recently, the development of LLMs, including GPT-3 [19], LLaMA [20] and Mixtral [21], has pushed the boundaries of what can be achieved with transformers, demonstrating unprecedented capabilities in understanding and generating human-like text. They also highlighted the limitations of LLMs in terms of biases (e.g., hallucinations [22]) and computational limitations due to their size.

RAG combines the strengths of retrieval methods and generative LLMs [14], bridging the gap between IR and text generation. In this approach, a retriever selects relevant documents or passages based on a query, and a generative model uses the retrieved information to produce a contextualised response

[23, 24]. By creating a dynamic and query-specific context, RAG enables LLMs to focus their attention on the most relevant information, improving accuracy and reducing hallucinations [25]. This method offers a powerful alternative to models based exclusively on static parameters.

2.3. Biomedical Q&A

Over the past twelve years, BioASQ evaluation campaigns [26] have enabled the development of various methods for BQA, which followed the evolution of IR and Answer Generation. Early approaches focused on extractive summarization techniques based on lexical matching, e.g., TF-IDF or LexRank. Over the years, participating teams started to use supervised and deep learning methods which outperformed previous works.

More recently, researchers have gained attention for transfer learning with models pre-trained on general domain of BQA datasets and fine-tuned on the BioASQ dataset [27]. Another step forward was the emergence of domain-specific LMs like PubMedBERT [8] which enabled to effectively encode biomedical entities and relational information.

With the emergence of LLMs, participating teams have naturally gained interest for RAG pipelines. They use sparse [28, 29, 30] or hybrid [31, 32] methods for the retrieval part. Most of them employ a re-ranking module to select more relevant articles. For answer generation, the proposed approaches explore different context creation strategies (e.g., ICL, snippets extraction), and different models tuning (e.g., prompt format, fine-tuning, parameter tuning). For more details on the approaches used on BQA tasks, we invite the reader to refer to the survey [7].

3. Method

As mentioned previously, this work aims to address the issue of context reshaping in a RAG pipeline applied to BQA. This section formalizes the problem and the method we propose to tackle it. An overview of our pipeline is illustrated in Figure 1.

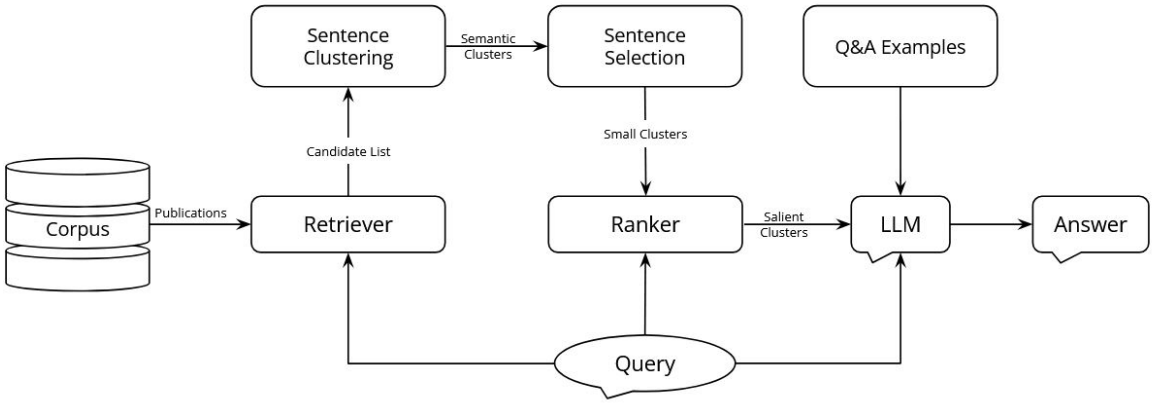

Figure 1: Overview of our overall Pipeline. 1000 publications are retrieved from PubMed Records. Their sentences are tokenized and grouped together in semantic clusters. The most important sentences are kept for each cluster, which are then ranked and selected by order of relevance to the query. This context, along with the query and examples of Q&A pairs are fed into a LLM to generate an answer.

3.1. Question Answering

This task can be defined as an answer generation depending on a context built from biomedical publications. Let q be a biomedical question expressed in natural language, and $D = \{d_1, d_2, ..., d_n\}$ a large set of biomedical publications. The system aims to generate an answer a by using a LLM and a

context C, which is extracted and potentially restructured from a subset $D' \subset D$. D' is obtained by running an IR module that should maximize Recall in order to contain the sought information.

3.2. Setting the Context

The documents of D' are decomposed into basic textual units, i.e., sentences. The obtained set $S = \{s_1, s_2, ..., s_k\}$ is composed of all the sentences extracted from D'. Each sentence s_i is encoded into a vector space using an encoder. The embedding of a sentence is produced by computing $mean\_pooling$ on the embedding of each token of the sentence. To simplify notations, we will only note:

$$E = \{SEncoder(s_i) \mid s_i \in S\},$$

$SEncoder(s_i)$ being the $mean\_pooling$ applied to the encoded tokens of s_i .

To guide the LLM attention during context processing, semantically close elements are grouped together. The intuition behind this idea is that a structured context will help the model "understand" the information given in input, instead of having information dispatched in S. The embeddings in E are grouped into clusters using cosine similarity, $E = \{E_1, E_2, ..., E_t\}$ where E_i denotes a cluster of embeddings. We note $T = \{T_1, T_2, ..., T_t\}$ the corresponding clusters of sentences, where each T_i is a group of sentences of a similar "topic":

$$T_i = \{s \in S \mid \forall e_1,\, e_2 \in E_i,\, cos\_sim(e_1,\, e_2) \geq threshold\}$$

For each cluster T_i, a ranking algorithm is applied to identify informative sentences, $T'_i \subset T_i$ and:

$$T'_i = sentenceRank(T_i,\, l),$$

where $sentenceRank$ is an implementation of a ranking method and l refers to the number of selected sentences.

Several works show that reordering documents can affect LLMs' performance and help them in the context processing [23]. To create our final context C, clusters in $T' = \{T'_1, T'_2, ..., T'_t\}$ are ranked based on their relevance to the query q. For a cluster T'_i, a cross-encoder produces a probability p_i of being relevant to q :

$$p_i = cross\_encoder(q,\, T'_i)$$

The most relevant clusters are then ranked to build C:

$$C = \{T'_{i_1}, T'_{i_2}, ..., T'_{i_c} \mid p_{i_1} \geq p_{i_2} \geq ... \geq p_{i_c}\},$$

where $c < t$ is the number of clusters to select.

3.3. Answer Generation

The context C, combined with instructions I, and the query q are fed into a LLM to generate the answer a:

$$a = LLM(I,\, q,\, C)$$

This methodology aims at generating highly contextualized and relevant answers to q by leveraging specialized documents while minimizing noise and irrelevant information.

4. Experiments

In this section, we present the datasets and metrics used to run our experiments. We also detail implementation settings for the retriever, the sentence selection, the topic ranking, and the answer generation modules.

4.1. Datasets

As shown in [7], there are few datasets directly addressing the specific task we tackle. We chose to work on the BioASQ-TaskB [33] dataset as it fits our specifications (see Section 1). BioASQ-TaskB is composed of two phases. Phase A aims at retrieving the 10 most relevant publications for a given query from the biomedical literature database PubMed[2], and extract their relevant snippets. Phase B focuses on answer extraction and generation by proposing an "exact answer" and an "ideal answer". "Exact answers" have a particular format depending on question type ("Yes/No", "Summary", "Factoid", "List"). "Ideal answers" are natural language texts that a biomedical expert could write to answer queries. To produce answers, participating teams are provided with the ground truth from Phase A, i.e, relevant articles and corresponding snippets. Since BioASQ 12, Phase A+ has been introduced. Its goal is the same than Phase B but without ground truth from Phase A. All BioASQ-TaskB data are manually annotated by biomedical experts, providing gold standards for various biomedical NLP tasks.

We isolated queries and their corresponding "ideal answers" from BioASQ's 11 and 12 campaigns, which enabled to evaluate our work on two distinct collections composed of 327 and 340 biomedical queries.

4.2. Metrics

BioASQ organizing team offers a manual evaluation of answers generated by participating systems. Each annotator gives a score out of 5 for the precision, recall, readability, and repetition criteria. ROUGE2 and ROUGE-SU4 (Recall, F1) scores [34] are also provided. Manual scores are computed only while the evaluation campaign is running. To evaluate our work and compare our models' performance with the methods proposed during the evaluation campaign, we have chosen to use ROUGE2 Recall, Precision, and F1. These metrics will be referred to as R2-R, R2-P, and R2-F1 respectively. However, there is an intrinsic limit to these lexical metrics when applied to text generation tasks. ROUGE2 evaluates the bi-gram overlap between a reference text and a candidate response. The score obtained by an answer semantically identical to the reference but using synonyms will obtained a very low score despite a correct answer. To evaluate our models, we have therefore used a metric based on semantic similarity, i.e., BERTScore [35]. On the one hand, we will be able to situate the performance of our approaches with R2 metrics, on the other hand we will have a more accurate idea of their performance with semantic similarities.

4.3. Retriever

We built a sparse retriever using Pyserini, an open-source Python library derived from Anserini which integrates multiple IR techniques. First, we indexed all MEDLINE citations except those for which the abstract was unavailable ($\approx$ 25M citations). For each query, we created a list of the thousand most relevant articles to answer it. This follows the observations of [36], showing that this architecture achieves a Recall@1000 greater than 90%. Considering the savings in ressources and computing time, we assert that this solution is suitable enough.

4.4. Sentence selection

The retriever makes it possible to find the publications that could include the context needed to respond to the query. The second step is to select the right information among all publications. We decided to

[2]https://pubmed.ncbi.nlm.nih.gov/

work at the sentence level to incorporate knowledge related to the query. We chose to compute an embedding of each sentence using an encoder-based model to enable a semantic comparison between them. We used the SentenceTransformer [37] library along with BioLinkBERT-large [9] to produce these embeddings. BioLinkBERT is a version of LinkBERT pre-trained on Biomedical corpora and achieving the best overall performance on the BLURB benchmark [38]. SentenceTransformer computes a sentence embedding by applying a mean pooling on the embeddings of the tokens composing this sentence.

We decided to group sentences by topic using a clustering method on the sentence embeddings. Since there were several thousand sentences to compare, we used the *community_detection* algorithm implemented by SentenceTransformer as it is designed to handle a large number of sentences. It computes the cosine similarity between embeddings to determine groups and incorporates several optimisations to manage large collections.

After semantically grouping together the sentences, it is needed to identify which sentences of each topic would compose the context. We implemented the TextRank algorithm and applied it to each cluster to identify their salient sentences (i.e., 4, 10, or 15 sentences per topic).

4.5. Ranker

In order to have the most precise context possible, it could be beneficial to choose the topics and eventually delete irrelevant ones for the given query. Furthermore, several studies have shown that the organization of the context can impact LLMs' performance in question answering tasks. Following our previous work on document ranking in a multi-stage retrieval pipeline [39], we draw an analogy between scientific publication ranking and topic ranking tasks. The former aims to rank documents by order of relevance to a query. We showed that changing the granularity of such documents and selecting relevant sentences among them instead of considering the whole document is beneficial. The topic ranking is globally the same task, differing only by the fact we work on clusters composed of semantically close sentences. We applied a BioLinkBERT cross-encoder fine-tuned on the BioASQ-TaskB dataset. This model computes a probability of relevance used to rank the topics. Once the ranked list of topics has been generated, we chose a fixed number to be used as context for the queries (i.e, the first 5, 10, or 15 topics depending on the experiments).

4.6. Answer Generation

We have seen in the previous sections how to establish a context for answering biomedical questions. To provide an answer based on a question and its context, we built an answer generation tool relying on LLaMA. We used the third release of this model in its 8B parameters version[3], called *llama3.1-8B* in the remainder of this paper. This model is open-weight and achieved SOTA results compared to LLMs of the same scale. Its architecture is optimized for high performance on Q&A tasks. Moreover, its powerful tokenizer enables to process a large number of input tokens, which is very important for ICL. We also chose this model to reduce computational costs and energy consumption compared to LLMs with higher number of parameters. To further reduce costs, we applied model quantization and used 4-bit precision for floating-point representation instead of 32-bit.

The prompts we used to parameterize the model are reported in Figures 2, 3, and 4. We also ran an experiment without context to evaluate the benefit of adding context. In these formulations, we first give a role to the system, then we explain the input that we give to the system, and finally we specify the task. We did not consider any prompt engineering optimization.

[3]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

System Prompt

You are a biomedical expert providing answers. I will give a question and several context texts about the question. Based on the context, give a short answer to the question.

User Prompt

QUESTION: **A biomedical query**
CONTEXTS: **Sentences extracted from biomedical literature**
ANSWER:

Figure 2: Prompt Specifications with Context Incorporation.

System Prompt

You are a biomedical expert providing answers. I will ask a question and your role is to give a short answer to the question.

User Prompt

QUESTION: **A biomedical query**
ANSWER:

Figure 3: Prompt Specifications without Context.

System Prompt

You are a biomedical expert providing answers. I will give a question and several context texts about the question. Based on the context, give a short answer to the question. Moreover I will give you 3 questions and their corresponding answers as examples.

User Prompt

EXAMPLES: **A set of 3 questions/answers dependent on the question type**
QUESTION: **A biomedical query**
CONTEXTS: **Sentences extracted from biomedical literature**
ANSWER:

Figure 4: Prompt Specifications with ICL and Context Incorporation.

5. Results

In this section, we present the experimental results obtained applying our approach on the two BQA datasets described in Section 4.1. We evaluated its performances by studying the impact of context incorporation and its reshaping. Then, we tested the effect of ICL by adding examples of (query, answer) pairs.

5.1. Influence of context texts

The aim of these first experiments is to show the effect of different types of context reshaping. We evaluate if few context text is enough for the LLM to perform well or if each piece of information needs to be repeated in order to be taken into account.

We developed three variants of the model to establish baselines. First, we generated answers using *llama3.1-8B* without incorporating any context. Next, we used *llama3.1-8B* on the same dataset but incorporating context by selecting 4 sentences per cluster and without applying any topic ranking. Finally, we extracted what we call the "Exact Context", which corresponds to the relevant snippets provided in BioASQ dataset. In a real-world scenario, such information is not available and this variant enables us to estimate the maximum scores achievable with this model configuration.

The scores obtained by these three variants on the BioASQ11 dataset are reported in Table 1. We observe that the basic system (without context) performs relatively well in terms of Recall but is very weak when it comes to Precision, whether semantic (BERT-P) or lexical (R2-P). The incorporation of context without ordering is undeniably beneficial, as it improves the basic system performance. However, we note that the highest improvements are primarily observed in the semantic metrics, indicating that while the LLM can leverage the context, it is not fully aligned with the vocabulary used by the annotators.

Table 1
Baselines on the BioASQ11 dataset

Context	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
None	28.86	14.50	16.50	58.52	31.50	35.50
Unranked Topics	30.60	14.78	16.73	69.90	60.77	64.40
Exact Context	37.02	26.76	26.70	73.43	67.68	69.70

5.2. Influence of context reshaping

We studied the impact of organizing the context. To do so, we added the cluster-ranking module to our previous experiments and generated answers while varying the parameters that define the context format, i.e., the number of clusters selected and the number of sentences per cluster. Since each cluster is associated with a topic present in the corpus, the objective here is to determine the required context size for generating accurate responses and to evaluate whether the LLM needs repeated information to effectively process it.

Table 2 presents the scores obtained for these experiments. First, we observe that, with 4 sentences per topic as in the previous experiments, selecting any number of clusters tends to decrease both lexical and semantic Recall scores. This was expected as we intentionally limited the amount of information retrieved. This loss is outweighed by the gain in Precision when selecting 10 clusters, as evidenced by the improvements in F1 scores. It appears that using too few or too many topics decrease the performance. We miss part of the information with few clusters, but adding too many introduces noise. This observation is in line with the fact that the ranking model is optimized to return a list of 10 relevant documents.

Afterwards, we studied the effect of increasing the number of sentences in each topic with fixed numbers of clusters. We generated answers with 5 or 10 clusters and for each configuration run the

experiment with 4, 10, and 15 sentences. We observe that each setup achieves higher scores as the number of sentences increases. In this case, we push less relevant topics further away from the query (regarding token distance in the sequence) without removing them. As a result, we give more weight to the relevant topics. Therefore, it seems wise to help the LLM focus its attention on the more relevant information without deleting less relevant ones.

Note that best scores for this set of experiments are achieved when using the parameters leading to the highest scores for each study (e.g., 15 sentences per clusters and 10 clusters). Moreover, we ran a *t-test* between this variant and the results obtained by the baseline labeled "Unranked Topics" in the previous section. The obtained p-values were lower than 0.05 on all metrics, meaning that all the improvements are significant.

Table 2
Effect of context structuring on an answer generation task on the BioASQ11 dataset

#sentences/cluster	#clusters	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
4	5	29.27	15.61	17.04	68.84	60.97	64.09
4	10	30.36	16.69	18.13	69.48	61.67	64.77
4	15	29.59	15.31	17.02	69.35	60.31	63.95
10	5	29.95	15.88	17.43	69.23	61.42	64.46
15	5	31.5	17.05	18.38	69.56	61.78	64.73
10	10	31.22	**17.56**	**18.77**	70.00	61.87	65.02
15	10	**31.52**	17.34	18.74	**70.43**	**62.43**	**65.54**

5.3. Influence of ICL

We decided to complete the incorporation of structured context by combining it with ICL. For each question type, we randomly extracted 3 examples of (question, answer) pairs from the BioASQ10 dataset. ICL is expected to help the LLM better understand how to structure its responses and potentially aligned itself on the vocabulary used by the annotators. Tables 3 and 4 show the scores obtained on the BioASQ11 and BioASQ12 datasets, respectively. We conducted experiments by varying the same parameters as in the previous section and using the prompt shown in Figure 4.

First, we observe that, for the same parameters, adding ICL consistently decreases performance on the two Recall measures: on average -5.17% on R2-R and -0.92% on BERT-R. This slight loss is more than compensated by higher gains in both Precision and F1 scores: on average +22.34% on R2-P and +4.32% on BERT-P. Table 5 shows the mean number of tokens in ground truth and in answers generated with 10 sentences and 10 clusters. Generated answers are much bigger than the gold standard and ICL tends to reduce answer length. This leads to retrieving slightly less information, but at the same time, the information returned is much more accurate. This phenomenon is observable on both datasets. The scores on lexical metrics are lower on BioASQ12 set. This can be explained by the fact that a new annotator was involved in its creation. Consequently, the model has no prior insight into the vocabulary used by this annotator. We ran a *t-test* between the best variant in Table 3 and our "Unranked Topics" baseline. We found that the p-value associated with BERT-R was higher than 0.05, meaning the loss is insignificant. All other p-values were lower than 0.05. The gains on Precision metrics are significant, but so is the loss on R2-R.

We compared our results (10 sentences and 10 clusters in Table 4) with other systems submitted in Phase A+ of the BioASQ12 challenge[4]. The best submissions in terms of R2-R (32.01 to 38.68 depending on the batch) have significantly lower Precision (R2-F1 ranging from 12.44 to 19.23) than our system (an average of 19.67 over the 4 batches). This indicates that our Recall-Precision trade-off is better. Moreover, the systems achieving the highest R2-F1 scores (25.03 to 28.62) exhibit a better trade-off but their corresponding Recall scores (R2-R ranging from 22.62 to 27.23) are lower than those

[4]https://participants-area.bioasq.org/results/12b/phaseAplus/

achieved by our system (average of 28.60). Considering that the top-performing runs used models with much more parameters (e.g., GPT-3.5, GPT-3.5 Turbo, GPT-4), employed fine-tuning techniques, and possibly leveraged metric-specific tuning (e.g., generated bigger answers to obtain a better Recall, used a translation module to optimize bi-grams overlap), we can conclude that our approach is both relevant and effective.

Table 3
Effect of ICL in BioASQ11

#sentences/cluster	#clusters	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
4	5	28.31	18.68	18.63	68.28	63.16	65.00
4	10	27.73	19.58	19.05	68.48	63.85	65.35
10	5	28.74	19.23	19.13	69.00	63.94	65.63
10	10	**29.75**	**23.05**	**21.28**	**69.22**	**65.61**	**66.58**

Table 4
Effect of ICL in BioASQ12

#sentences/cluster	#clusters	R2-R	R2-P	R2-F1	BERT-R	BERT-P	BERT-F1
4	5	26.91	17.70	18.10	68.44	62.33	64.69
4	10	27.03	18.87	18.77	68.66	63.53	65.37
10	5	27.64	**19.75**	18.38	69.56	61.78	64.73
10	10	**28.60**	19.67	**19.61**	**70.10**	**64.41**	**66.50**

Table 5
Mean number of tokens

Generation tool	BioASQ11	BioASQ12
No ICL	97.0	110.5
With ICL	74.6	80.24
Gold Standard	42.1	50.3

6. Conclusion

In this article, we presented several approaches to incorporate biomedical knowledge into a LLM. We showed that answers generated with contextualized prompts has more influence on Precision than on Recall. Moreover, improvements on semantic metrics are more important than on lexical ones, meaning the generated answers are not easily aligned with a given vocabulary.

Ranking clusters enhances the scores under specific conditions. It is essential to select enough clusters to capture relevant information, but retrieving too many can introduce noise and degrade performance. In addition, it seems beneficial to increase the number of sentences per cluster. This helps the LLM focusing its attention on relevant information by pushing away less relevant information without deleting it. Finally, we show that integrating ICL to a RAG pipeline, despite a slight loss on Recall, enables major improvements in terms of Precision and F1 scores. The comparison with some of the top-performing models from the BioASQ challenge shows that our approach achieves competitive results.

Future work will be dedicated to study alternative ways to incorporate context for answer generation, e.g., using biomedical knowledge-bases to structure knowledge in semantic predications (subject-predicate-object triples) [40, 41]. Further investigations into optimizing context selection could improve both answer quality and readability in real-world biomedical applications. Finally, it would be wise to integrate citations directly in the answers so that readers can easily validate the generated information.

Our work aligns with the challenge of efficiently extracting and structuring biomedical knowledge from vast amounts of scientific literature. In fields like metabolomics, where researchers must analyze large amounts of publications to interpret metabolic signatures, automated methods could significantly assist knowledge retrieval. Existing tools like FORUM[5] facilitate bibliographic exploration by linking metabolites to biomedical concepts, but they remain limited in handling large-scale textual data. By refining context selection and integrating structured knowledge representations, our approach could help improving literature-based discovery in metabolomics and beyond.

7. Declaration on Generative AI

During the preparation of this work, the authors used Generative AI tools in order to: Grammar and spelling check, Text Translation, Improve writing style. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

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

[2] A. Radford, K. Narasimhan, Improving language understanding by generative pre-training, 2018. URL: https://api.semanticscholar.org/CorpusID:49313245.

[3] A. Yates, R. Nogueira, J. Lin, Pretrained transformers for text ranking: BERT and beyond, in: G. Kondrak, K. Bontcheva, D. Gillick (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials, Association for Computational Linguistics, Online, 2021, pp. 1–4. URL: https://aclanthology.org/2021.naacl-tutorials.1/. doi:10.18653/v1/2021.naacl-tutorials.1.

[4] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, J. rong Wen, Large language models for information retrieval: A survey, ArXiv abs/2308.07107 (2023). URL: https://api.semanticscholar.org/CorpusID:260887838.

[5] Q. Zhang, K. Ding, T. Lv, X. Wang, Q. Yin, Y. Zhang, J. Yu, Y. Wang, X. Li, Z. Xiang, et al., Scientific large language models: A survey on biological & chemical domains, ACM Computing Surveys (2024).

[6] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261/. doi:10.18653/v1/2020.findings-emnlp.261.

[7] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, S. Yu, Biomedical question answering: a survey of approaches and challenges, ACM Computing Surveys (CSUR) 55 (2022) 1–36.

[8] R. Tinn, H. Cheng, Y. Gu, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Fine-tuning large neural language models for biomedical natural language processing, CoRR abs/2112.07869 (2021). URL: https://arxiv.org/abs/2112.07869. arXiv:2112.07869.

[9] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, 2022. arXiv:2203.15827.

[10] K. r. Kanakarajan, B. Kundumani, M. Sankarasubbu, BioELECTRA:pretrained biomedical text encoder using discriminators, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 143–154. URL: https://aclanthology.org/2021.bionlp-1.16. doi:10.18653/v1/2021.bionlp-1.16.

[5] https://forum-webapp.semantic-metabolomics.fr

[11] Q. Dong, Y. Liu, S. Cheng, S. Wang, Z. Cheng, S. Niu, D. Yin, Incorporating explicit knowledge in pre-trained language models for passage re-ranking, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1490–1501. URL: https://doi.org/10.1145/3477495.3531997. doi:10.1145/3477495.3531997.

[12] J. Tan, J. Hu, S. Dong, Incorporating entity-level knowledge in pretrained language model for biomedical dense retrieval, Computers in Biology and Medicine 166 (2023) 107535.

[13] Q. Xie, P. Tiwari, S. Ananiadou, Knowledge-enhanced graph topic transformer for explainable biomedical text summarization, IEEE Journal of Biomedical and Health Informatics 28 (2024) 1836–1847. doi:10.1109/JBHI.2023.3308064.

[14] 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.

[15] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1107–1128. URL: https://aclanthology.org/2024.emnlp-main.64/. doi:10.18653/v1/2024.emnlp-main.64.

[16] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, X. Cheng, Semantic models for the first-stage retrieval: A comprehensive review, ACM Trans. Inf. Syst. 40 (2022). URL: https://doi.org/10.1145/3486250. doi:10.1145/3486250.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[20] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[21] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al., Mixtral of experts, arXiv preprint arXiv:2401.04088 (2024).

[22] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[23] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, F. Silvestri, The power of noise: Redefining retrieval for rag systems, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 719–729. URL: https://doi.org/10.1145/3626772.3657834. doi:10.1145/3626772.3657834.

[24] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, Y. Shoham, In-context retrieval-augmented language models, Transactions of the Association for Computational Linguistics 11 (2023) 1316–1331.

[25] O. Ayala, P. Bechard, Reducing hallucination in structured outputs via retrieval-augmented generation, in: Y. Yang, A. Davani, A. Sil, A. Kumar (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 228–238. URL: https://aclanthology.org/2024.naacl-industry.19/. doi:10.18653/v1/2024.naacl-industry.19.

[26] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková,

A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.
[27] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.
[28] S. Ateia, U. Kruschwitz, Can open-source llms compete with commercial models? exploring the few-shot performance of current GPT models in biomedical tasks, 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, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 78–98. URL: https://ceur-ws.org/Vol-3740/paper-07.pdf.
[29] Y. Gao, L. Zong, Y. Li, Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation, in: CLEF (Working Notes), 2024, pp. 117–129. URL: https://ceur-ws.org/Vol-3740/paper-10.pdf.
[30] J. H. Merker, A. Bondarenko, M. Hagen, A. Viehweger, Mibi at bioasq 2024: retrieval-augmented generation for answering biomedical questions, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, volume 3740, 2024, pp. 176–187.
[31] T. Almeida, R. A. Jonker, J. Reis, J. R. Almeida, S. Matos, Bit. ua at bioasq 12: From retrieval to answer generation, CLEF Working Notes (2024).
[32] D. N. Panou, A. C. Dimopoulos, M. Reczko, Farming open llms for biomedical question answering, 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, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 188–196. URL: https://ceur-ws.org/Vol-3740/paper-17.pdf.
[33] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 12b and Synergy12 in CLEF2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.
[34] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[35] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
[36] T. Almeida, R. A. A. Jonker, R. Poudel, J. M. Silva, S. Matos, Bit. ua at bioasq 11b: Two-stage ir with synthetic training and zero-shot answer generation., in: CLEF (Working Notes), 2023, pp. 37–59.
[37] 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.
[38] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthcare 3 (2021). URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.
[39] M. Lesavourey, G. Hubert, Enhancing Biomedical Document Ranking with Domain Knowledge Incorporation in a Multi-Stage Retrieval Approach., in: 12th BioASQ Workshop at CLEF 2024, volume 3740, Grenoble, France, 2024. URL: https://hal.science/hal-04744454.
[40] G. Agrawal, T. Kumarage, Z. Alghamdi, H. Liu, Can knowledge graphs reduce hallucinations in LLMs? : A survey, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3947–3960. URL: https://aclanthology.org/2024.naacl-long.219/. doi:10.18653/v1/2024.naacl-long.219.
[41] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T. C. Rindflesch, Semmeddb: a pubmed-scale

repository of biomedical semantic predications, Bioinformatics 28 (2012) 3158–3160.