How do you know that? Teaching Generative Language Models to Reference Answers to Biomedical Questions

Anonymous ACL submission

Abstract

Large language models (LLMs) have recently become the leading source of answers for users' questions online. Despite their ability to offer eloquent answers, their accuracy and reliability can pose a significant challenge. This is especially true for sensitive domains such as biomedicine, where there is a higher need for factually correct answers. This paper introduces a biomedical retrieval-augmented generation (RAG) system designed to enhance the reliability of generated responses. The system is based on a fine-tuned LLM for the referenced question-answering, where retrieved relevant abstracts from PubMed are passed to LLM's context as input through a prompt. Its output is an answer based on PubMed abstracts, where each statement is referenced accordingly, allowing the users to verify the answer. Our retrieval system achieves an absolute improvement of 23% compared to the PubMed search engine. Based on the manual evaluation, our LLM component achieves better results than GPT-4 Turbo in referencing relevant abstracts, increasing recall by up to 18%. We make the dataset used to fine-tune the model and the fine-tuned models based on Mistral-7B-instruct-v0.1 and v0.2 available.

1 Introduction

The idea of automated referencing dates back to 1970 when (Garfield, 1970) proposed an automatic system where a computer evaluates the appropriateness of references within an article. With the emergence of large generative language models (LLMs), numerous systems are being developed to answer specific questions, supported by relevant references (Huang and Chang, 2024; Menick et al., 2022; Yang et al., 2023). Generative LLMs can produce answers that appear coherent, confident and articulate. However, the information conveyed may not be correct or verifiable. Furthermore, the limited internal knowledge of generative LLMs can hinder their ability to deliver factually accurate answers, particularly within specialized fields (Gravel et al., 2023; Zheng et al., 2023). This issue is notably concerning in the biomedical domain, where accurate and factual answers are critical. The scientific community have recognized the dangers of factually incorrect or nonsensical information and has been reluctant to utilize these models to their potential. Providing an opportunity for scientists to obtain correct and verifiable answers to questions is an opportunity to increase scientific productivity and its impact. Moreover, privacy, sovereignty and security concerns in pharma and biomedicine often necessitate building systems where all components are controllable (e.g. deployed in-house), to avoid reliance on third-party APIs such as OpenAI[1].

Incorporating domain-specific external knowledge beyond LLMs data is essential for mitigating hallucinations in LLMs. The retrieval-augmented generation (RAG) approach, which integrates the generative capabilities of an LLM with a specialized retrieval system enhances the model's accuracy and relevance by grounding its responses in verified information.

In this paper, we present a biomedical RAG system consisting of a hybrid search based on PubMed[2] and fine-tuned generative model for referenced question-answering. We make both the models and the dataset used to fine-tune the models publicly available.

The remainder of this paper is organized as follows: Section 2 provides a review of related work on reliability and verifiability of the LLM generated content and the approaches to generating texts with references. Section 3 describes the design of the IR and generative components. We evaluate the components in Section 4, first individually and then jointly. We end the paper with conclusions

[1] https://openai.com
[2] https://pubmed.ncbi.nlm.nih.gov/download/

and some future work remarks in Section 5.

2 Related work

Generative LLMs, such as GPT and similar architectures, have enabled question-answering (QA) tasks across various domains, including medicine. The current state of these models is characterized by several challenges, particularly regarding the verifiability and reliability of the information they generate. By evaluating ChatGPT responses and references in the medical domain, (Gravel et al., 2023) found that 69% of generated references were fabricated, while professionals rated the answers at a median quality of 60%. Similarly, when (Liu et al., 2023) conducted manual evaluations of four prominent generative search engines Bing Chat, NeevaAI, perplexity.ai, and YouChat, they found that while the responses from these engines were fluent and seemingly informative, only 51.5% of sentences generated by these engines were fully supported by their citations, and merely 74.5% of citations accurately supported the statements they were linked to.

In general, there are two approaches to generating text with references (Huang and Chang, 2024). The first one assumes training LLMs to produce references from parametric knowledge (information internalized from the training data). The second one assumes producing references from non-parametric knowledge (content retrieved from external sources).

The first approach, integrating citations directly from LLM's parametric knowledge, poses a significant technical challenge. Unlike search engines and IR systems that rely on indices for data retrieval, LLMs encode information into hidden representations during training, lacking a direct index. Therefore, referencing the sources of information becomes intricate. Despite these challenges, approaches have been suggested to train LLMs to include references using source identifiers (Taylor et al., 2022). However, these methods exhibit certain limitations, including citation inaccuracies and being restricted to academic citations.

The second approach, known as retrieval-augmented generation (RAG), combines generative LLMs with IR systems to form a hybrid system (Lewis et al., 2020). Here, the model is trained to recognize instances requiring citations, and the IR system retrieves suitable sources to provide context to the LLM. As a result, the LLM incorporates these sources as citations into its outputs, improving the credibility and accuracy of responses. While pre-trained and fine-tuned LLMs rely solely on their parametric knowledge, RAG integrates a customized external knowledge base without additional training, thus reducing hallucinations. Moreover, annotators often perceive RAG-enhanced answers to be more factual and specific compared to those from fine-tuned models (Lewis et al., 2020).

3 Method

The RAG system we propose in this paper is designed to perform referenced question-answering (QA) in the biomedical domain. It consists of two main components. The IR component, based on hybrid semantic and lexical search, retrieves relevant PubMed abstracts and provides a context for fine-tuned generative LLM. The final system output is an answer to the user query, which contains a reference for each of the claims extracted from the relevant abstracts. The overview of the system architecture can be seen in Figure 1.

3.1 Information Retrieval Component

Our IR component uses data from PubMed database[3] containing citations and biomedical literature from several literature resources. The IR system integrates both sparse vectors (lexical index) and dense vectors (semantic index), enabling lexical and semantic search, and a hybrid combination of the two.

For the lexical retrieval, based on BM25, we use the OpenSearch[4] to create an index for PubMed articles, by concatenation of title and abstract as an indexed field. Also, we add authors' names, publication dates, and journal names as metadata for filtering.

For semantic retrieval, based on dense vectors, we use the Qdrant[5] vector database. Qdrant allowed the usage of memory mapping of vectors to hard drive, reducing memory (RAM) requirements of the system. In order to optimize semantic search retrieval time, we used 8-bit quantized embeddings, with the option to use full embeddings for rescoring the results.

We use the Hierarchical Navigable Small World (HNSW) indexing technique for Approximate Nearest Neighbors with dot product metrics to perform vector comparisons (Malkov and Yashunin,

[3] https://pubmed.ncbi.nlm.nih.gov/download/
[4] https://opensearch.org/
[5] https://qdrant.tech/

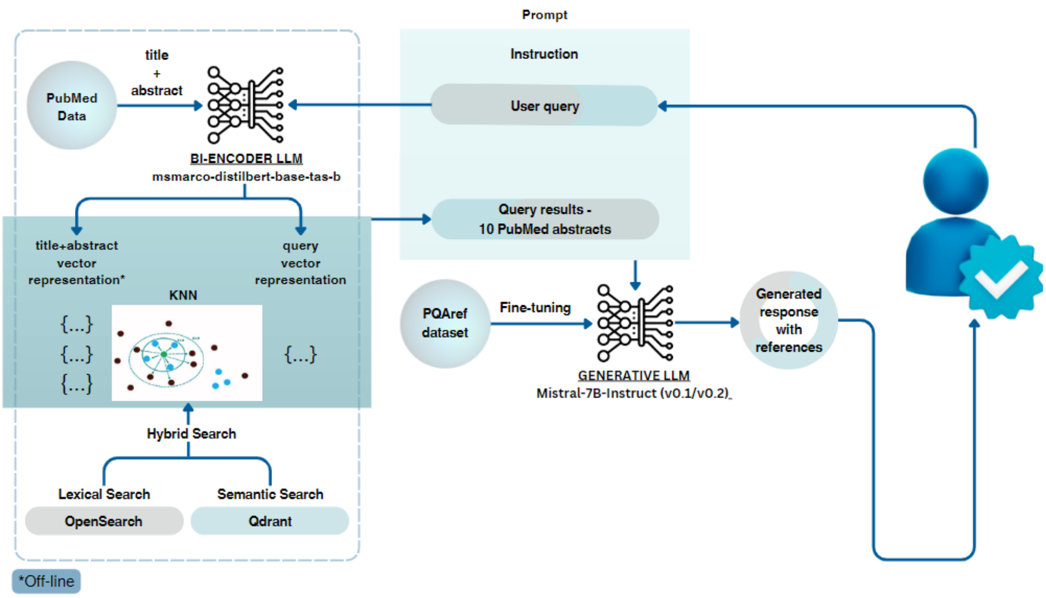

Figure 1: Architecture of our RAG system.

2018). To create vector embeddings we use a bi-encoder sentence transformer model pre-trained on the MSMarco dataset (Hofstätter et al., 2021), which, at the time of indexing, had the best performance on Passage Retrieval Task[6].

Among a corpus of 36,797,469 abstracts, 11,308,679 were found to be empty and thus omitted from the index. Before generating embeddings for semantic search, it was ascertained that the average number of tokens within the dataset's title and abstract concatenation was 650. Given that the maximum input size of the model employed for embedding creation is 512 tokens, abstracts exceeding this threshold were subdivided into segments, each containing no more than 512 tokens, and indexed separately. The split was made at the end of the sentence before the 512th token.

Hybrid search is in our case combination of lexical and semantic IR components. To utilize a hybrid search, we normalized scores from these two IR methods to scales ranging from 0 to 1. The scores from each of the search methods are then multiplied with the importance weights for each of the methods (based on the experimental evidence, the best weight combination was 0.3 for lexical and 0.7 for semantic component) This allows the identification of both direct matches and greatly improves the ability to discover semantically related phrases and text segments, even in the absence of exact textual matches.

[6] https://www.sbert.net/docs/pretrained-models/msmarco-v3.html

3.2 Generative Component

3.2.1 Dataset

We created a custom dataset to fine-tune an LLM for the task of question-answering (QA) with references. The dataset consists of 9,075 questions, each question is provided with 10 relevant abstracts (along with titles and PMIDs) and referenced answers to the questions based on the provided abstracts.

The questions were randomly selected from the PubMedQA dataset (Jin et al., 2019). The most relevant abstracts for each of these questions were retrieved from the PubMed repository using a combination of entity and free text search. To create the answers based on the retrieved abstracts, we used GPT-4 Turbo, currently the number one model on the Chatbot Arena leaderboard, a crowdsourced open platform for LLM evaluation (Chiang et al., 2024). The prompt we used to instruct GPT-4 Turbo to use references (PMIDs) was as follows:

> Answer the question using relevant abstracts provided, up to 300 words. Reference the statements with the provided abstract_id in brackets next to the statement.

To ensure the completeness of answers, GPT-4 Turbo was further instructed to continue generating if there is more content to generate. The answers were then automatically checked for completeness and incomplete final sentences were removed, which finally led to the size of answers ranging from 69 to 1221 tokens. In a small number of cases (25 questions) there is no direct answer in the abstracts so the answer does not contain any references. The total input length in the dataset (question + abstracts + answer) ranges from 1686 to 6987 tokens.

We name this dataset PQAref and plan to make it freely available.

3.2.2 Fine-tuning the models

The generative component of our system is based on the Mistral-7B model. Despite having fewer parameters, Mistral-7B shows superior performance over larger models such as Llama 2 13B across all evaluated benchmarks and Llama 1 34B in reasoning benchmarks, maths, and code generation (Jiang et al., 2023). Compared to its 0.1 version, Mistral-7B v0.2 introduced an expanded context window (32K to the previous 8K) and several other adjustments (rope-theta = 1e6, no sliding-window attention) contributing to more accurate and consistent outputs, improved efficiency, and adaptability to many different tasks (Anakin.ai, 2024). For our task, we use instruction-tuned versions of Mistral-7B, both v0.1[7] and v0.2[8]. We will refer to those two models as M1 and M2 in the rest of the paper.

Both instruction-tuned versions were fine-tuned for the task of referenced QA using the QLoRA methodology (Dettmers et al.), allowing us to fine-tune the models on a single DGX NVIDIA A100-40GB GPU in ∼32 hours. The parameters we used for both models were standard loss, rank of 64, alpha of 16, and LoRA dropout of 0.1, resulting in 27,262,976 trainable parameters in both cases. Both models were fine-tuned over 2 epochs, using a batch size of 1. The PQAref dataset split we used was 80:10:10, with most inputs in the size range of 4000 to 6000 tokens in all three splits (see Figure2).

We make the QLoRA adapters for both models available on ANONYMIZED.

To obtain the answers from both fine-tuned models, we used the following prompt:

> Respond to the Instruction using only the information provided in the relevant abstracts in "'Abstracts"' below.
> {instruction}
> Answer:

The *instruction* consists of the user query and 10 retrieved abstracts. We use default inference parameters for both models, except setting the repetition_penalty to 1.1 and varying the values of max_new_tokens. Despite adding the limit to the answers through the max_new_tokens parameter or through trying to add a limit to the prompt (e.g. "Answer in at most 300 words."), both models continuously generated an arbitrary number of tokens. The same behavior was noticed in GPT-4 Turbo during the creation of the PQAref dataset. The token limitation, primarily imposed due to the prolonged inference time for higher values (e.g. 1000) would only lead to interrupted answers. Finally, the limit was set to 1225, to match the longest complete answer length in the training dataset. An example of all three models' answers to an instruction from the test set can be seen in Appendix A1.

4 Results

4.1 Evaluation of IR Component

To evaluate our IR system, we utilized the BioASQ dataset (BioASQ team, 2024). The BioASQ dataset is designed for tasks that help drive advancements in biomedical information retrieval and QA. It includes 5049 questions along with corresponding gold-standard answers, relevant document snippets, and the PubMed IDs (PMIDs) of articles that are relevant to each question.

We compared the PMIDs retrieved by our system against the gold-standard PMIDs provided in the BioASQ dataset. This comparison was quantified using the precision metric, measuring the proportion of relevant identifiers retrieved by our system out of the total PMIDs retrieved. We evaluate precision using 10 retrieved documents (P@10) and mean average precision for 10 retrieved documents (MAP@10). The evaluation of the retrieval component is done using: (1) only lexical, (2) only semantic, and (3) a combination of the two. Additionally, we experimented with different weights for the lexical and semantic combinations.

For the Lexical search, we experimented with stopword removal from the query and obtained better results compared to lexical without stopwords removal as shown in Table 1.

[7] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1`

[8] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

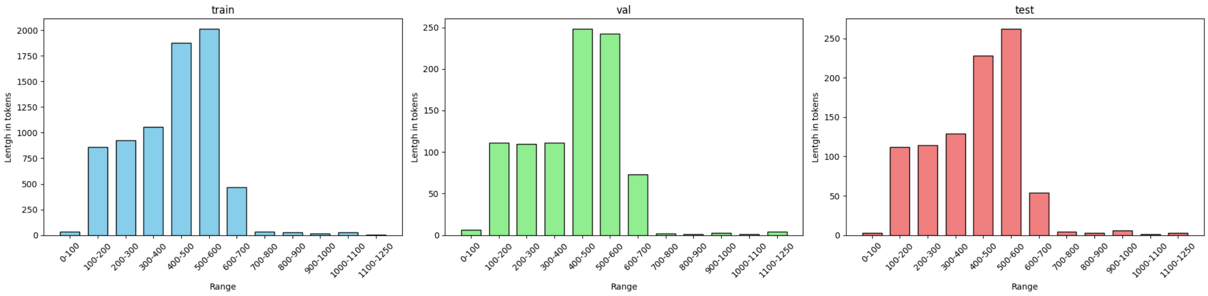

Figure 2: Distribution of answer length across train, val and test splits.

For semantic search, we experimented with three approaches: semantic search with full embedding, semantic search with compressed embeddings (using 8-bit quantization), and semantic search using compressed embeddings with rescoring (using full embeddings for rescoring).

Semantic search with full embedding had an average response time of 30 seconds, making it inefficient and unusable for real world applications.

For semantic search with rescoring, we used compressed embeddings to retrieve 100 results, then rescored the top 10 using full-size embeddings. This method improved precision by 0.3% and was only 52 milliseconds slower than the approach without rescoring (see 1st and 2nd rows in Table 1). Given the minimal additional time required, we tested the various weight combinations of hybrid search incorporating semantic search with restoring. Parallel execution of semantic and lexical search further contributes to the time efficacy of the system (as shown in Table 1, reducing average execution time from 489ms to 442ms).

Table 1: Our IR and PubMed search engine performance evaluation on the BioASQ dataset.

	P@10	MAP@10	time [ms]
1. Semantic without rescore	14.0%	25.7%	245
2. Semantic with rescore	14.4%	26.0%	297
3. Hybrid with rescore (lex. 0.1 sem. 0.9)	24.7%	32.5%	442
4. Hybrid with rescore (lex. 0.2 sem. 0.8)	24.7%	32.5%	442
5. Hybrid with rescore (lex. 0.3 sem. 0.7)	24.7%	32.5%	442
6. Hybrid with rescore (lex. 0.4 sem. 0.6)	24.7%	32.6%	442
7. Hybrid with rescore (lex. 0.5 sem. 0.5)	25.2%	41.0%	442
8. Hybrid with rescore (lex. 0.6 sem. 0.4)	30.7%	42.0%	442
9. Hybrid with rescore (lex. 0.7 sem. 0.3)	**30.8%**	**42.5%**	**442**
10. Hybrid with rescore (lex. 0.8 sem. 0.2)	30.8%	42.5%	442
11. Hybrid with rescore (lex. 0.9 sem. 0.1)	30.8%	42.6%	442
12. Lexical with stopwords removal	28.7%	41.1%	189
13. Lexical without stopwords removal	28.3%	40.1%	189
PubMed without Mesh Terms	9.2%	15.3%	698
PubMed with Mesh Terms	12.0%	19.1%	742

From the experiments detailed in Table 1, it is evident that the performance of semantic search alone is suboptimal, with notable enhancements observed upon integration with lexical search. The initial improvement is noted with the hybrid search employing a 0.1 lexical search weight, followed by a second significant enhancement achieved with a 0.6 lexical search weight (yielding absolute improvements of 10.3% and 16.3% respectively). Increasing the lexical search weight beyond 0.6 does not yield noticeably different outcomes. Assigning a weight of 1 to lexical search in hybrid search excludes semantic search, effectively reducing the system to pure lexical search, which produces worse results.

As the subsequent generative component does not account for the order of retrieved documents, we employ the P@10 metric to determine the most effective combination of parameters for hybrid search. After evaluating various configurations, we identified the optimal parameters for hybrid search: a lexical search weight of 0.7 and a semantic component weight of 0.3. Consequently, we choose these parameter values, as shown in row 9, to conduct a hybrid search in our system.

Additionally, we evaluated the performances of PubMed search engine on BioASQ dataset and got a precision of P@10 12% and MAP@10 19.1%.

4.2 Evaluation of the Generative Component

We conducted automated and manual evaluations for the referenced QA task. This involved analyzing the total number of references, distinguishing between general and relevant references, checking the correctness of IDs, and comparing the number of relevant references to irrelevant ones.

Automated evaluation. The number of referenced abstracts in generated answers within PQAref test set (containing 908 examples) can be seen in Table 2. What can be observed is that 1 reference per answer is most common for GPT-4 Turbo (241 answers), while M1 and M2 have the highest number of answers with 3 references (185 cases for M1 and 178 for M2). The average number of references per answer of all three models is similar, revolving around 4, with M2 offering the highest number of references per answer (4.2). In the entire test set, comprising 908 examples with a total of 9080 abstracts (10 abstracts per example), M1 and M2 did not reference any abstracts in 8 and 5 answers, respectively. In each of these instances, their conclusion was that none of the abstracts were relevant, demonstrating their proficiency in task execution.

Table 2: Number of referenced abstracts per model on the PQAref test set.

N	GPT-4 Turbo	M1	M2
0	2	8	5
1	**241**	86	105
2	76	138	112
3	128	**185**	**178**
4	126	172	169
5	119	117	124
6	87	72	75
7	45	66	34
8	29	27	34
9	31	22	23
10	24	15	49
Sum of referenced abstracts	3,464	3,648	3,816
% of 9080 abstracts	38.15	40.18	42.03
Avg no of references per answer	3.81	4.01	4.20

To measure the relevance of the referenced abstracts, we evaluated whether the models referenced at least the most relevant abstract for each question. Our dataset contains questions from PubMedQA, which in a number of cases originate from actual PubMed abstract titles. This means that during retrieval, the article whose title matches the question is very likely to be retrieved as relevant. In our test split, this indeed is the case in 823 out of 908 inputs. We decided to take such abstracts as the most relevant ones for those 823 inputs, which allowed us to automatically measure the number of times the models referenced that particular abstract. Table 3 presents the number of missed and referenced relevant abstracts using this tactic. GPT-4 Turbo missed the most relevant abstract in only one case, suggesting it is a good referencing role model. M2 missed the relevant abstract in 10 examples, while M1 missed it in 29 examples. Overall, both models do reference the most relevant abstract in most cases (96.5% and 98.8% respectively).

We also evaluated whether all the IDs in the models' answers matched the PMIDs of context-provided abstracts to verify none of them were hallucinated. GPT-4 Turbo's answers contained no hallucinated IDs. However, both M1 and M2 produced hallucinated IDs, with a notable discrepancy. M1 produced 79 hallucinated IDs, while M2 produced only 3. The hallucinated IDs differ from the actual IDs by one or two digits. Upon manual inspection of the answer content and referenced IDs, we found that M1 tended to blend information from various abstracts, whereas M2 strictly utilized information solely from the corresponding abstract. This suggests that M2 exclusively hallucinated some of the digits from the existing abstract ID, but not the content. This behavior remains consistent across different temperature values of the model.

Table 3: The number of missed and referenced relevant abstracts of 823 abstracts across the models.

	GPT-4 Turbo	M1	M2
Relevant missed	1 (0.1%)	29 (3.5%)	10 (1.2%)
Relevant referenced	822 (99.9%)	794 (96.5%)	813 (98.8%)

Manual evaluation. To perform manual evaluation, we extracted 10 random examples from the PQAref test set. We then manually assessed the relevance of each of the abstracts in the examples. We generally distinguished between two types of abstracts: relevant and irrelevant. The abstracts we considered *relevant* were the ones that covered all the specific aspects of the question and thus provided direct answers. Among them, the abstract whose title matched the question (similarly as mentioned for 823 examples in automatic evaluation) we defined as *the most relevant*. On the other hand, we identified two types of *irrelevant* abstracts. The first type includes abstracts that miss the main topic of the question (e.g. discuss heart failure instead of knee problems) which are considered *completely irrelevant* abstracts. The other type discusses a more general topic and thus does not cover all the aspects of the question, we considered *partially irrelevant*. This group could also be observed as the one that contains additional information but does not provide the direct answer to the question. It is crucial to recognize that the severity of mistakes differs between the two types. If the model references a completely irrelevant abstract that is a clear mistake, however, if it references a partially irrelevant abstract, whether it is wrong may depend on the other references in the answer. If the answer also contains the reference that gives a direct answer to the question (relevant abstract), this could be considered additional information. If this is not the case, the model may have missed the main point. For example, for the question "Could the central part of the partograph, the cervicograph, be improved?", the abstracts marked as relevant are those that specifically discuss the improvement of the cervicograph. Documents that focus their attention on the partograph without directly mentioning cervicograph are considered irrelevant.

Finally, we examined if the three models referenced the relevant abstracts. For these 10 qualitatively observed examples, all the models correctly referenced the most relevant abstracts and none of the models referenced completely irrelevant abstracts. The general tendency of all three models was to provide additional information by referencing partially irrelevant abstracts. In several situations, the models seemed to filter the abstracts based on their understanding of a term used in the question, thus excluding the abstracts that use a different phrasing or an extended meaning of the term (e.g. donation taken to refer only to organ, tissue or bone marrow donation and not to cell and blood donation).

We also conducted a quantitative analysis to examine how well they identified all the relevant abstracts. To overcome variations in the number of relevant abstracts per document and document-specific characteristics, we considered all 100 abstracts, 10 for each of 10 questions, collectively.

Of these 100 abstracts, the evaluators identified 42 relevant and 58 irrelevant abstracts. We prioritized and calculated recall for relevant abstracts for each model, as our primary concern is their ability to correctly identify and reference relevant abstracts. M1 exhibited the highest recall of 0.76, followed by M2 with 0.67, while GPT-4 Turbo showed the lowest recall of 0.62, as can be seen in Table 4. These findings suggest that, based on the analysis of these 10 manually reviewed documents, M1 outperforms the other two models in terms of referencing abstracts deemed relevant by evaluators.

Table 4: Recall values for relevant abstracts on 10 examples from the PQAref test set and same 10 questions with abstracts retrieved with our IR system.

	GPT-4 Turbo	M1	M2
PQAref	0.62	0.76	0.67
IR	0.46	0.64	0.58

4.3 System evaluation

In this section, we provide the preliminary joint evaluation of our system: the IR component (based on hybrid lexical and semantic search) and the generative component using the outputs of our IR and fine-tuned Mistral models.

We manually evaluated the IR output on the same 10 PQAref questions we chose for the evaluation of the generative component in Section 4.2. To retrieve the relevant abstract from indexed PubMed articles, we utilized the best-performing hybrid search parameter combination from Section 4.1 and retrieved 10 abstracts for each question. After manually determining the abstract relevance, we obtained 50% P@10. This metric underscores the effectiveness of our IR component in locating documents for query responses. The fact that IR evaluation on BioASQ reached the best performance of P@10 30.8% with the same combination of weights for hybrid search as manual evaluation on PQAref, further corroborates the results obtained in manual evaluation conducted on the PQAref dataset.

We then used the same prompts as in Section 4.2 for GPT-4 Turbo, as well as M1 and M2, to generate referenced answers based on the retrieved documents. We further computed the recall values for the relevant abstracts in the 10 generated answers. It is noticeable that the model that showed the highest recall of 64% is M1 (Table 4). This model cites a greater number of abstracts that contain the relevant answers compared to other models. From Table 2, we can also observe that the model with most citations is M2. However, it shows a slightly lower recall (58%) because it has fewer citations of abstracts that provide direct answers to the questions. Nonetheless, since the IR component consistently finds documents related to the topic, we give preference to M2's answers since it includes more additional citations, offering more elaborate answers on the same topics. Here, GPT-4 Turbo had the lowest recall of 46%.

5 Conclusions and future work

In this paper, we provide an overview of biomedical generative search with answers grounded in PubMed and referenced claims. Our aim was to develop a system capable of generating accurate and verifiable answers to biomedical questions while maintaining user sovereignty. We achieved this by fine-tuning open-source models and creating a solution independent of any third-party APIs. Starting with our IR component, we discovered that employing a combination of lexical and semantic searches yields the highest precision score. Our system demonstrates an absolute improvement of 23.4% in the MAP@10 measure compared to the PubMed search engine. Through separate evaluations, we found that lexical search alone outperforms semantic search. However, integrating both approaches is advantageous for identifying instances lacking exact term matches, where semantic search contributes significantly. To enhance semantic search performance in IR, one future direction is to fine-tune these models on domain-specific data. This approach aims to improve the quality of embeddings in the biomedical domain, enabling them to encode domain-specific knowledge better, enhance contextual understanding, and ultimately improve IR performance.

Overall, fine-tuned Mistral 7B Instruct models performed comparatively to GPT-4 Turbo in terms of the task of referenced QA. Based on the evaluation of the whole PQAref test set, M1 and M2 referenced a high number of the most relevant abstracts, with M2 performing better than M1 by 2.3% and just 1.1% worse than GPT-4 Turbo. As a general trend, M2 includes more information in its answers.

Both models showed hallucinations when generating IDs of references. However, M2 performed better than M1 with only 3 mismatches in ciphers (0.08% of the referenced abstracts), compared to 79 mismatches in M1 (2.16% of the referenced abstracts), while using correct information from the corresponding abstract. Nevertheless, this point needs further attention. Exchanging the IDs with numerals (1-10) for each abstract during fine-tuning could potentially solve this issue and is something we plan to try in the next iteration of the dataset and training.

Based on the manual evaluation of 10 examples from PQAref test set and the same 10 questions with abstracts retrieved with our IR system and their respective recall values for relevant abstracts, both M1 and M2 had higher values than GPT-4 Turbo. While recall values for relevant abstracts were higher for M1 on our example sets, M2 had a close-enough performance, differing by only 9% on PQAref and 6% using our IR system. Overall, our RAG system achieves increased recall values up to 18% compared to GPT-4 Turbo.

It is worth noting that we have achieved a better performance than GPT-4 Turbo with much smaller, open-source fine-tuned models, albeit on a small test set. Having an automated evaluation method of reference quality would allow us to make a more comprehensive comparison of our models to models such as GPT-4 Turbo. To supplement manual evaluation and speed up the evaluation process, (Gao et al., 2023) created a benchmark for automatic evaluation of citation quality and proposed an entailment model as a method for automatic quality check of references. We intend to further check the quality of answers on a larger sample using the same method. At this point, M2 gives more reliable answers, but the decision on which model we will use in our final pipeline will be made after this evaluation.

References

Anakin.ai. 2024. Mistral 7b v0.2 base model, the new open source llm king is here. https://anakin.ai/blog/mistral-7b-v0-2-base-model/. Accessed: 2024-04-16.

BioASQ team. 2024. Bioasq12 challenge dataset. [Accessed 26 April 2024].

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Eugene Garfield. 1970. Can citation indexing be automated? *Essays of an Information Scientist*, 1:84–90.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: Limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of SIGIR 2021*.

Jie Huang and Kevin Chen-Chuan Chang. 2024. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Nan Yang, Tao Ge, Liang Wang, Binxing Jiao, Daxin Jiang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Inference with reference: Lossless acceleration of large language models. *arXiv preprint arXiv:2403.12077*.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does ChatGPT fall short in providing truthful answers? *arXiv preprint arXiv:2304.10513*.

Appendix A: Example answers of generative LLMs

Table A1: Models' answers to the question "Does leisure time physical activity in early pregnancy protect against pre-eclampsia?"

GPT-4 Turbo
The relationship between leisure time physical activity (LTPA) in early pregnancy and the risk of pre-eclampsia is not clearly established, with studies showing mixed results. Some studies suggest that there may be a protective effect of LTPA against pre-eclampsia, while others do not find a significant association or suggest an increased risk with high levels of activity. A prospective cohort study in Denmark found that moderate levels of physical activity (1-270 minutes/week) in the first trimester had no statistically significant association with the risk of pre-eclampsia, but high levels of activity (exceeding 270 minutes/week) were associated with an increased risk of severe pre-eclampsia (PUBMED:19055653). Similarly, a population-based prospective cohort study in Amsterdam reported that the amount of time or intensity of physical activity in leisure time was not associated with a difference in risk of pre-eclampsia or gestational hypertension (PUBMED:19916877). In contrast, a study from the New Haven, Connecticut area suggested that women who engaged in any regular LTPA during early pregnancy were at decreased risk of pre-eclampsia (PUBMED:15466498). Another study found a tendency towards a lower risk of pre-eclampsia in women with the highest degree of physical activity during leisure time, especially in overweight women, although no significant associations were found (PUBMED:20121498). A systematic review examining the association between LTPA before and/or during pregnancy and the risk of pre-eclampsia found mixed results, with some studies suggesting that vigorous-intensity LTPA may reduce the risk of pre-eclampsia, while one study reported an elevated risk with high amounts of LTPA (PUBMED:23836014). Another case-control study found that women who performed regular LTPA during the first 20 weeks of pregnancy had a reduced risk of pre-eclampsia and gestational hypertension (PUBMED:2592903). In summary, the evidence is inconclusive regarding the protective effect of LTPA in early pregnancy against pre-eclampsia. While some studies suggest a potential benefit, especially with moderate activity levels, others indicate no effect or an increased risk with high levels of activity. It is important to consider the intensity and duration of LTPA when evaluating its impact on pre-eclampsia risk.
M1
Based on the provided abstracts, leisure time physical activity in early pregnancy does not appear to protect against pre-eclampsia. A prospective cohort study in Denmark found that the two highest levels of physical activity were associated with increased risk of severe pre-eclampsia compared with the nonexercising group, suggesting that leisure time physical activity exceeding 270 minutes/week in first trimester may increase the risk of severe pre-eclampsia (PUBMED:19055653). Another study in Amsterdam, The Netherlands, did not find any association between physical activity in leisure time early in pregnancy and the incidence of preeclampsia or gestational hypertension (PUBMED:19916877). Similarly, a study in New Haven, Connecticut, found that women who engaged in any regular leisure time physical activity, regardless of caloric expenditure, were at decreased risk of preeclampsia, but there was no indication of a protective effect of workplace activity, leisure-time physical activity, or unemployment (PUBMED:15466498). These findings suggest that while certain forms of physical activity may have a protective effect on preeclampsia, the overall evidence does not support a clear protective effect of leisure time physical activity in early pregnancy against this condition. Therefore, based on the available abstracts, leisure time physical activity in early pregnancy does not protect against pre-eclampsia.
M2
Based on the provided abstracts, there is conflicting evidence regarding whether leisure time physical activity (LTPA) in early pregnancy protects against pre-eclampsia. Some studies suggest that LTPA may offer protection against pre-eclampsia, while others do not find a significant association or even indicate that excessive LTPA could increase the risk of severe pre-eclampsia. For instance, one study found that women who engaged in any regular LTPA regardless of caloric expenditure were at decreased risk of preeclampsia, although another study did not find a significant association between the amount of time or intensity of physical activity in leisure time and the risk of preeclampsia or gestational hypertension (PUBMED:15466498; PUBMED:19916877). Another study suggested that women who performed regular LTPA had a reduced risk of pre-eclampsia and gestational hypertension, with the relative risks decreasing as the average time spent in LTPA increased (PUBMED:2592903). However, another study documented an increased risk of severe pre-eclampsia with higher levels of LTPA (PUBMED:19055653). Similarly, a systematic review reported that high-intensity LTPA before and/or during pregnancy or more than 4 hours per week of LTPA may reduce the risk of pre-eclampsia, but also noted that one study found an elevated risk of severe pre-eclampsia with high amounts of LTPA (PUBMED:23836014). Furthermore, a case-control study indicated that women who performed regular LTPA had a reduced risk of pre-eclampsia and gestational hypertension, with the relative risks decreasing as the average time spent in LTPA increased (PUBMED:26910608). A systematic review and meta-analysis of RCTs and cohort studies supported the promotion of LTPA in pregnancy as a strategy to improve maternal and child health, suggesting that participation in LTPA was associated with lower weight gain during pregnancy, lower likelihood of gestational diabetes mellitus (GDM), and lower risk of preterm delivery (PUBMED:27282925). In conclusion, while some studies suggest that LTPA in early pregnancy may protect against pre-eclampsia, particularly at moderate intensities, the evidence is not consistent across all studies. Further research is needed to clarify the relationship between LTPA and pre-eclampsia risk, taking into account the intensity and duration of physical activity, as well as individual differences in susceptibility to the condition. Additionally, the potential influence of other factors such as prepregnancy BMI and socioeconomic status should be considered (PUBMED:32093248).