Exploring the Few-Shot Performance of Low-Cost Proprietary Models in the 2024 TREC BioGen Track

Samy Ateia and Udo Kruschwitz
Information Science, *University of Regensburg*, Regensburg, Germany
{Samy.Ateia, Udo.Kruschwitz}@sprachlit.uni-regensburg.de

***Abstract*—For the 2024 TREC Biomedical Generative Retrieval (BioGen) Track, we evaluated proprietary low-cost large language models (LLMs) in few-shot and zero-shot settings for biomedical question answering. Building upon our prior competitive approach from the CLEF 2024 BioASQ challenge, we adapted our methods to the BioGen task. We reused few-shot examples from BioASQ and generated additional ones from the test set for the BioGen specific answer format, by using an LLM judge to select examples. Our approach involved query expansion, BM25-based retrieval using Elasticsearch, snippet extraction, reranking, and answer generation both with and without 10-shot learning and additional relevant context from Wikipedia. The results are in line with our findings at BioASQ, indicating that additional Wikipedia context did not improve the results, while 10-shot learning did. An interactive reference implementation that showcases Google's Gemini-1.5-flash performance with 3-shot learning is available online[1] and the source code of this demo is available on GitHub[2].**

***Index Terms*—Domain-Specific IR, Large Language Models, Few-Shot Learning, TREC BioGen Challenge, Query Expansion, Question Answering, Professional Search**

I. Introduction

The TREC BioGen Track aims to advance biomedical information retrieval by evaluating systems that generate accurate and comprehensive answers to biomedical queries, with a focus on correct reference attribution to mitigate false statements by LLMs [1].

In our previous work in the CLEF BioASQ challenge[3] [2], we explored the performance of commercial and open source LLMs in a similar retrieval augmented generation (RAG) setting for biomedical question answering, using available training data for fine-tuning and few-shot learning [3]. In the BioASQ challenge, the systems are evaluated separately in document retrieval, snippet extraction, and answer generation for different answer formats (yes/no, factoid, list, ideal).

For the TREC BioGen challenge, the participating systems are evaluated on a short (maximum 150 words) textual answer with inline citations to PubMed IDs. Precision and recall of the retrieved documents is indirectly evaluated by the PubMed IDs cited in this answer text. This answer format is comparable to the "ideal" answer format at the BioASQ challenge, but it extends it by requiring accurate inline citations.

Our main goal for participating was to verify our results from the BioASQ challenge, where our approach was competitive, winning multiple first and second spots in different rounds of the challenge[4]. We also wanted to test the newest proprietary low-cost commercial models and compare their performance against each other and the other systems that participated in the Track.

II. Methods

A. Data and Resources

We indexed the titles and abstracts of all papers in the 2023 annual baseline of PubMed[5] in Elasticsearch using the built-in default English analyzer[6]. The LLMs used are proprietary models: OpenAI's GPT4o-Mini and Google's Gemini-1.5-flash-001 [4]. For query expansion, snippet extraction, and snippet reranking, we reused few-shot examples taken from our participation at the 2024 CLEF BioASQ challenge. Because the answer format for BioGen differs from the answer format at BioASQ, we generated few-shot examples from the test set and used an LLM judge to select examples for our few-shot learning approach.

B. Retrieving Relevant Wikipedia Context

To retrieve relevant context from Wikipedia, the models were prompted with the question and narrative and instructed to generate a list of existing Wikipedia article titles with helpful information to answer this question. These articles were then retrieved, and the models were again prompted to extract and summarize the most relevant information to answer the question. This summary was used in some runs as additional context information when generating queries, extracting and ranking snippets or answering questions.

C. Query Expansion and Transformation

For each query, we performed query expansion and transformation using the LLMs to generate boolean queries compatible with Elasticsearch's `query_string` syntax[7]. Additional

[1] https://bioragent.samyateia.de/

[2] https://github.com/SamyAteia/BioRAGent

[3] https://www.bioasq.org/

[4] https://www.bioasq.org/participate/twelfth-challenge-winners

[5] https://pubmed.ncbi.nlm.nih.gov/download/\#annual-baseline

[6] https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-lang-analyzer.html\#english-analyzer

[7] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html\#query-string-syntax

context from relevant Wikipedia articles was incorporated in some runs to see if it improves query expansion.

D. Retrieval Process

1) **Initial Retrieval**: We conducted BM25-based retrieval on the title and abstract fields in our PubMed index.
2) **Query Refinement**: If zero results were returned, an optional query refinement step was performed.

E. Snippet Extraction and Reranking

The LLMs were prompted to extract snippets from the top 50 retrieved documents. The snippets were then reranked by the LLM based on the relevance to the question, and the top 20 snippets were selected for answer generation.

F. Answer Generation

Answers were generated using the LLMs in either zero-shot or ten-shot settings. Few-shot examples for this task were created from the test set with GPT-4o and rated by an LLM judge (GPT4o-Mini) prompted with instructions taken from the BioGen evaluation guideline[8].

G. Interactive System Implementation

An interactive system implementation is available online at https://bioragent.samyateia.de/, which utilizes 3-shot learning with Gemini-1.5-flash-001 [5]. The ”Answer with Citations” field corresponds to the answer format required in the BioGen challenge, while the ”Answer” fields showcase the ”ideal” answer format from BioASQ.

III. Results

A. Runs Submitted

We submitted the six runs listed below. Zero-shot or 0s indicates runs where no examples were given in the corresponding prompts to the LLM. Ten-shot or 10s signifies that the run used 10 examples in the prompts, and the ”-wiki” suffix designates that additional context from Wikipedia was added to the prompt.

- **Zero-Shot GPT4o-Mini (0s 4o-mini)**
- **Zero-Shot Gemini-Flash (0s Gemini)**
- **Ten-Shot GPT4o-Mini (10s 4o-mini)**
- **Ten-Shot Gemini-Flash (10s Gemini)**
- **Ten-Shot GPT4o-Mini-Wiki (10s 4o-mini-wiki)**
- **Ten-Shot Gemini-Flash-Wiki (10s Gemini-wiki)**

B. Evaluation Metrics

Our runs were evaluated by the BioGen Track organizers based on the following metrics, focusing on both answer quality and citation correctness:

- **Answer Quality**: Accuracy, Completeness (Recall), Precision, Redundancy Score, Harmfulness Score.
- **Citation Quality**: Citation Coverage, Citation Support Rate, Citation Contradict Rate.
- **Document Relevancy**: Recall and Precision.

C. Answer Quality

The answer **accuracy** metric evaluates the proportion of acceptable answers produced by each run out of the total 65 questions. An acceptable answer is one that addresses the question at least partially. According to the information provided by the organizers, the expected accuracy is 90% and above. Table I presents the answer accuracy for our runs.

TABLE I
Answer Accuracy

Run Name	Acceptable	Accuracy (%)
10s 4o-mini	65	100
0s 4o-mini	64	98.46
10s Gemini	62	95.38
10s 4o-mini-wiki	61	93.85
10s Gemini-wiki	61	93.85
0s Gemini	56	86.15
Mean (All Teams)	**60.13**	**92.51**
Max (All Teams)	**65**	**100**
Min (All Teams)	**39**	**60**

The *10s 4o-mini* run achieved an accuracy of 100%, providing acceptable answers to all questions. The *0s 4o-mini* run followed closely with 98.46% accuracy. The runs with additional Wikipedia context, *10s 4o-mini-wiki*, and *10s Gemini-wiki* performed worse than their non-augmented 10-shot counterparts. The *0s Gemini* run had the lowest accuracy among our submissions at 86.15%. Compared to the mean accuracy across all teams (92.51%), our runs performed well, with the *10s 4o-mini* matching the maximum accuracy of 100%.

Additional quality metrics assess precision, redundancy, and harmfulness at the answer level.

Precision measures the proportion of required assertions correctly included in the answers. **Redundancy Score** quantifies the proportion of unnecessary sentences in the answers. **Harmfulness Score** assesses the proportion of potentially harmful sentences in the answers. Table II summarizes these metrics. The *10s 4o-mini* run had the highest precision at 83.93% and a redundancy score of 12.90%, with a harmfulness score of 0%. Other runs had precision values between 68.40% and 81.98%, redundancy scores from 12.59% to 17.13%, and low harmfulness scores across the board. Compared to the mean precision across all teams (77.39%), our top runs performed above average, but also had above average (lower is better) redundancy scores.

D. Citation Quality Metrics

Citation quality metrics evaluate the correctness and relevance of references in the answers.

Citation Coverage measures how well answer sentences are supported by appropriate citations. **Citation Support Rate** assesses alignment between system-predicted citations and human-judged supportive citations. **Citation Contradict Rate** penalizes inclusion of citations that contradict the answer statements. As shown in Table III, the *10s Gemini* run achieved the highest citation coverage at 72.94% and a support rate

[8] https://dmice.ohsu.edu/trec-biogen/evaluation.html

TABLE II
ANSWER QUALITY METRICS

Run Name	Precision (%)	Redundancy (%)	Harmfulness (%)
10s 4o-mini	83.93	12.90	0
0s 4o-mini	81.98	12.59	0.22
10s Gemini	75.88	15.21	0.48
10s 4o-mini-wiki	75.62	15.37	0
10s Gemini-wiki	74.07	17.13	0
0s Gemini	68.40	13.09	0
Mean (All Teams)	**77.39**	**11.44**	**0.25**
Max (All Teams)	**90.68**	**26.41**	**1.54**
Min (All Teams)	**52.79**	**3.51**	**0**

of 64.02%, with a contradict rate of 0.26%. Other runs had citation coverage between 60.61% and 66.35% and support rates from 56.59% to 68.63%. Contradict rates remained low across all runs. Compared to the mean citation coverage (63.56%) and support rate (56.74%) across all teams, most of our runs performed above average.

TABLE III
CITATION QUALITY METRICS

Run Name	Coverage (%)	Support (%)	Contradict (%)
10s 4o-mini	66.35	68.63	1.88
10s Gemini	72.94	64.02	0.26
10s 4o-mini-wiki	60.61	65.77	0.19
10s Gemini-wiki	65.89	56.59	1.26
0s 4o-mini	64.13	60.56	1.55
0s Gemini	63.06	58.13	1.35
Mean (All Teams)	**63.56**	**56.74**	**2.23**
Max (All Teams)	**92.21**	**79.97**	**6.13**
Min (All Teams)	**0**	**0**	**0**

E. Document Relevancy Metrics

Document relevancy metrics assess the system's ability to retrieve relevant documents.

Recall measures the proportion of all relevant documents retrieved. **Precision** evaluates the proportion of retrieved documents that are relevant. Table IV presents these metrics. The *10s 4o-mini* run achieved the highest recall among our submissions at 9.06% and a precision of 73.62%. Other runs had recall values ranging from 6.51% to 8.89% and precision between 60.22% and 70.64%. Compared to the mean recall (7.79%) and precision (65.85%) across all teams, our runs performed slightly above average in both precision and recall.

IV. DISCUSSION

Our experiments indicate that the 10-Shot GPT4o-Mini run achieved the highest scores among all our submitted runs. The incorporation of additional context from Wikipedia did not improve performance, but instead led to worse performing runs. This is in line with our previous findings at BioASQ, where additional Wikipedia context did not improve results.

TABLE IV
DOCUMENT RELEVANCY METRICS

Run Name	Recall (%)	Precision (%)
10s 4o-mini	9.06	73.62
10s Gemini	8.89	69.23
10s 4o-mini-wiki	8.40	69.72
10s Gemini-wiki	8.73	60.22
0s 4o-mini	8.29	70.64
0s Gemini	6.51	61.22
Mean (All Teams)	**7.79**	**65.85**
Max (All Teams)	**24.12**	**90.04**
Min (All Teams)	**0**	**0**

A. Few-Shot Learning with Limited Data

By creating few-shot examples from the test set and utilizing an LLM judge for selection, we demonstrated that few-shot learning can lead to improvements in RAG-based question answering even if there is no ground truth data available to take few-shot examples from. A limitation of our approach is that we used a more capable LLM (GPT-4o) to create the examples. It is therefore not clear if the improvements just stem from the 10 examples solved by a better model, or actually few-shot learning effects.

B. Model Comparison

For zero-shot learning, OpenAI's GPT-4o-mini performed better than Google's Gemini-1.5-flash-001 in most metrics. Only the citation contradiction score was slightly better for Gemini-1.5-flash-001. These models are the cheapest offerings from both vendors, promising the best trade-off between generation quality, speed, and price.

C. Document Relevancy

Recall rates were relatively low across all runs, indicating room for improvement in the retrieval component, but still higher than the average score for all teams, except for Gemini-1.5-flash-001 in the zero-shot setting. The document retrieval results were also in line with our results at the BioASQ challenge, where our system didn't secure top spots in the retrieval task, but performed above average. Our query expansion approach with normal BM25 based retrieval might not reach the performance of vector-based or hybrid retrieval systems yet, but it offers the advantage that the used semantic knowledge of the model can be explicitly made visible to the user improving transparency and controllability which can be important in professional search use-cases [6].

V. CONCLUSION

We explored the performance of two proprietary LLMs, GPT4o-Mini and Gemini-1.5-flash-001 in zero-shot and few-shot settings for the TREC BioGen challenge. Our findings suggest that OpenAI's GPT4o-Mini performs better in terms of answer accuracy and precision than Google's Gemini-1.5-flash-001. Given the evaluation results and the best of our runs performing above average, we have the impression that these high speed low-cost models are viable options to build interactive domain-specific RAG systems. Overall, the results are consistent with our findings at BioASQ, where few-shot learning improved performance, and additional Wikipedia context did not. Future work may involve enhancing the retrieval component, exploring additional LLMs, and refining few-shot learning strategies. The most interesting topic might be to explore when and how information and knowledge from both search results or additional knowledge bases improves or hurts the performance of generative LLMs in domain-specific tasks.

Acknowledgments

We thank the TREC BioGen organizers for providing the challenge framework and resources.

References

[1] D. Gupta, D. Demner-Fushman, W. Hersh, S. Bedrick, and K. Roberts, “Overview of TREC 2024 Biomedical Generative Retrieval (BioGen) Track,” *arXiv preprint arXiv:2411.18069*, 2024.

[2] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, and G. Paliouras, “Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and N. Ferro, Eds., 2024.

[3] S. Ateia and U. Kruschwitz, “Can Open-Source LLMs Compete with Commercial Models? Exploring the Few-Shot Performance of Current GPT Models in Biomedical Tasks,” in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, ser. CEUR Workshop Proceedings, G. Faggioli, N. Ferro, P. Galuscáková, and A. G. S. de Herrera, Eds., vol. 3740. CEUR-WS.org, 2024, pp. 78–98. [Online]. Available: https://ceur-ws.org/Vol-3740/paper-07.pdf

[4] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arXiv preprint arXiv:2403.05530*, 2024.

[5] S. Ateia and U. Kruschwitz, “BioRAGent: A Retrieval-Augmented Generation System for Showcasing Generative Query Expansion and Domain-Specific Search for Scientific Q&A,” *arXiv preprint arXiv:2412.12358*, 2024.

[6] A. MacFarlane, T. Russell-Rose, and F. Shokraneh, “Search strategy formulation for systematic reviews: Issues, challenges and opportunities,” *Intelligent Systems with Applications*, vol. 15, p. 200091, 2022.