Generating Long-Form Answers to Biomedical Questions: Effectiveness and Efficiency

Jan Bakker and **Jaap Kamps**
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands
`{j.bakker, kamps}@uva.nl`

Abstract

In this paper, we report on the University of Amsterdam's participation in the TREC 2025 BioGen Track (Gupta et al., 2025). The goal is to generate answers to biomedical questions that are grounded with appropriate citations. First, we investigated the zero-shot generalization of PISCO, an efficient open-domain question answering method, to this task. Second, we compared the use of small and large Llama models. Our findings reveal a trade-off between answer quality and generation time. Therefore, we emphasize the need for better approaches to combining LLM reliability with efficiency.

1 Introduction

This paper reports on the University of Amsterdam's participation in the TREC 2025 Biomedical Generative Retrieval (BioGen) Track. We participated in task B on reference attribution. Given a corpus of PubMed abstracts, the goal is to generate answers to biomedical questions using LLMs. Importantly, each output sentence should contain up to three citations to PubMed abstracts that support the answer. The task of reference attribution was introduced in 2024 as a means to mitigate the generation of false statements by LLMs (Gupta et al., 2024). This year, 30 new questions were made available (Gupta et al., 2025).

Since the goal is to generate answers that are grounded with appropriate citations, we must provide citable abstracts as input to an LLM. These abstracts can be retrieved in the Retrieval stage of a Retrieval-Augmented Generation (RAG) pipeline. In the Augmented Generation stage, the LLM should then generate an answer based on the abstracts that were estimated to be the most relevant. In general, larger models tend to generate better answers. However, including multiple contexts (such as abstracts) in the input of large language models substantially slows down their decoding time, affecting the time a user has to wait for an answer. Therefore, we conducted two experiments with regard to LLM efficiency.

First, we investigated the zero-shot generalization of PISCO (Louis et al., 2025), an efficient question answering method, to the reference attribution task. PISCO is largely based on COCOM (Rau et al., 2025) and both methods effectively reduce long contexts to a handful of context embeddings, speeding up generation time by a large margin. However, since the corresponding models were trained on open-domain question answering datasets, they were trained to generate very short outputs. Our use of PISCO rather than COCOM is motivated by the inclusion of a prompt in the input of PISCO models. By altering this prompt, we were able to investigate the zero-shot generalization of these models to long-form biomedical question answering with reference attribution.

Second, we compared the answer generation capacities of a relatively small Llama model (8 billion parameters) with that of a larger Llama model (70 billion parameters). Our goal was to investigate to what extent LLM size really affects answer quality. Since our initial findings indicated that PISCO generalizes poorly to the task at hand, we prioritized our Llama runs for manual evaluation. The judgements for these runs confirm that in general, the larger model tends to generate better answers.

The remainder of this paper is structured as follows. In Section 2, we describe our experimental setup. The results of our experiments are then described in three phases in Section 3. Finally, we end with a discussion and conclusions in Section 4.

Run	**Generator LLM**
UAmsterdam_bergen_llama-70b	meta-llama/Llama-3.3-70B-Instruct (4-bit quantized)
UAmsterdam_bergen_llama-8b	meta-llama/Meta-Llama-3.1-8B-Instruct
UAmsterdam_bergen_pisco-llama	naver/pisco-llama (8B)
UAmsterdam_bergen_mistral-7b	mistralai/Mistral-7B-Instruct-v0.2
UAmsterdam_bergen_pisco-mistral	naver/pisco-mistral (7B)

Table 1: TREC 2025 BioGen Task B Submissions

Retriever	naver/splade-v3
Top-50 reranker	naver/trecdl22-crossencoder-debertav3

Table 2: Retrieval setup for all runs

2 Experimental design

We participated in task B of the TREC 2025 BioGen Track and submitted five runs as shown in Table 1. The retrieval setup was kept the same for all runs and is shown in Table 2. We used the BERGEN library (Rau et al., 2024) to perform our experiments. Compared to the Ragnarök framework used in the TREC 2024 RAG Track (Pradeep et al., 2025), the BERGEN library places more emphasis on Augmented Generation compared to Retrieval.

Retrieval We employed a SPLADE sparse encoder (Formal et al., 2021) and reranked the top-50 results for each input question with a DeBERTa (He et al., 2021) cross-encoder trained on MSMARCO v1 (Bajaj et al., 2018). This is the same setup that Rau et al. (2025) and Louis et al. (2025) used to retrieve passages from Wikipedia. In this case, we retrieved abstracts from the PubMed 2025 baseline. For those few abstracts that are longer than 512 tokens, retrieval was performed on a truncated version. The top-10 reranked abstracts per input question were ultimately passed on to the next stage.

Augmented Generation We conducted runs with five LLMs that generated answers to all 30 input questions based on the top-10 retrieval results per question. Two GPUs were used for llama-70b and one GPU otherwise. The batch size was set to 4 and the maximum number of tokens to generate was set to 350. The system prompt was kept the same for all runs and is shown in Figure 1; it is a slightly modified version of the example prompt included in the starter kit.[1] A shortened example of a user prompt is also shown in Figure 1.

For models that do not use context compression (llama-70b, llama-8b, mistral-7b), each abstract was simply tokenized together with the rest of the user prompt. For models that do use context compression (pisco-llama, pisco-mistral), each tokenized abstract was instead passed through the PISCO model together with a set of 8 memory tokens. The final hidden states that correspond to the memory tokens were then extracted to form the context embeddings for that abstract. These embeddings, with separator tokens appended, were ultimately included in the tokenized user prompt in place of the abstracts. Both PISCO models were trained to compress Wikipedia passages of 128 tokens into 8 memory tokens, effectively reducing their size by a factor of 16. Meanwhile, PubMed abstracts are twice as long on average, but vary in length.

Evaluation Because the PISCO models were trained to generate the same answers as their uncompressed counterparts (on a vast array of open-domain questions), they learned to effectively compress contexts (Wikipedia passages) and leverage the resulting context embeddings to answer open-domain questions. In this paper, we investigate the zero-shot generalization of these models to long-form biomedical question answering with reference attribution. To this end, we compared the quality of the answers generated by the PISCO models (pisco-llama, pisco-mistral) to that of their uncompressed

---

[1] `https://github.com/trec-biogen/starter-kit-2025`

System

Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly at the end of each sentence. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [2][4][7]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents.

User

Question: what is axonal injury on my mri brain report; is it a concern?

[1]: Prognostic Significance of Magnetic Resonance ... cases with severe head trauma.
[2]: Diffuse axonal injury: a case ... in the evaluation of DAI.
[3]: The comparison of the value ... DAI are of microscopic nature.
[4]: [Diffuse axonal injury manifested as ... diffuse axonal injury are discussed.
[5]: A Case Report Emphasizing an ... enhances their quality of life.
[6]: Diagnostic and Prognostic Capability of ... capability of various MRI sequences.
[7]: Diffuse axonal injury at ultra-high ... of ultra-high field (UHF) MRI.
[8]: [Predictive value of early MRI ... sensitive for detecting DAI lesions.
[9]: Diffuse axonal injury has a ... to severe traumatic brain injury.
[10]: Extensive traumatic axonal injury of ... known about TAI by violence.

Answer:

Figure 1: System prompt and user prompt template. A shortened example of a filled-in user prompt is shown in gray.

counterparts (llama-8b and mistral-8b). We also measured the average response time per LLM so that we could compare their efficiency. Furthermore, we included a 4-bit quantized llama-70b model in our experiments, so that we could assess the effect of LLM size on response quality and generation time.

3 Results

In this section, we present our results in three phases. First, we present an initial surface-level analysis of the quality of the generated answers. Second, we cover the answer generation times per language model. Third, we present and discuss the official scores for the manually and automatically evaluated answers.

3.1 Qualitative analysis

We present here an initial surface-level qualitative analysis of the generated answers, based on the example outputs shown in Figure 2. To begin with, we observe that the quantized llama-70b model appears to have successfully followed the instruction in the system prompt. Its answer to the illustrative question in Figure 1 seems to be coherent, complete, and grounded with citations. In contrast, the response generated by llama-8b is detailed rather than concise and ends abruptly because the maximum generation length was reached, yet its answer appears to be fluent and grounded with citations, too. Similar observations can be made about the other outputs generated by these two models.

Moreover, the answers generated by llama-8b and mistral-7b sometimes come across as summaries of the topic based on the provided abstracts, rather than focused answers to the given question. Compared to llama-8b, mistral-7b generated answers that are more concise, but frequently appended the titles of the references to their answer until the maximum generation length was reached.

Both PISCO models generated responses that are too short and tend to end mid-sentence, such as the one in Figure 2. This is clearly a result of the way these models were trained to respond with the name of a single entity. The main difference between the outputs of both models is that pisco-llama often did not attribute its references, whereas pisco-mistral did.

Llama-70b

Axonal injury, also known as diffuse axonal injury (DAI), is a type of brain injury characterized by extensive lesions in the white matter tracts of the brain, often caused by traumatic events such as car accidents, falls, or sports-related incidents [1][2][3]. It is a severe and frequently life-altering form of traumatic brain injury that can lead to unconsciousness, persistent vegetative state, and profound neurologic deficits [1][5]. The injury occurs due to shear and tensile forces that disrupt axons throughout the brain, impairing vital neural connections [5]. Axonal injury can be detected using magnetic resonance imaging (MRI), which can show lesions in the periventricular white matter, corpus callosum, and on the borderline of the white and grey matter [3][4]. The diagnosis of axonal injuries is difficult, as the majority of lesions found in DAI are of microscopic nature [3]. However, newer MRI sequences and techniques, such as multidimensional MRI, have shown promise in detecting and characterizing axonal injury [7][9]. If you have axonal injury on your MRI brain report, it is a concern and should be discussed with your doctor to determine the best course of action and treatment [6][8].

Llama-8b

Axonal injury on an MRI brain report refers to damage to the white matter tracts in the brain, which can occur due to shearing forces during traumatic events such as car accidents, falls, or sports-related incidents [3][5]. This type of injury is characterized by the widespread disruption of axons, leading to multifocal axonal damage and impairing vital neural connections [5]. Diffuse axonal injury (DAI) is one of the most common and devastating types of traumatic brain injury, accounting for about half of all cases with severe head trauma [1][2]. The diagnosis of axonal injuries is challenging, as the majority of lesions found in DAI are of microscopic nature [3]. However, magnetic resonance imaging (MRI) has been gaining popularity as an adjunctive imaging method in patients with DAI, and various MRI sequences have been used to evaluate DAI [2][6]. Studies have shown that MRI can detect DAI lesions, which are typically located in the periventricular white matter, corpus callosum, and on the borderline of the white and grey matter, subventricularly [3][4]. The prognosis of axonal injury varies depending on the severity of the injury and the effectiveness of treatment. Research suggests that patients with DAI may experience physical and psychosocial disabilities, and the prognostic value of DAI is still a matter of debate [8]. However, early MRI findings have been shown to have predictive value for neurocognitive and psychiatric outcomes in patients with severe traumatic brain injury [8]. In some cases, axonal injury may not be visible on conventional MRI, but can be detected using diffusion tensor imaging (DTI) [10]. Multidimensional MRI, which

Pisco-llama

Axonal injury on an MRI brain report, specifically in the context of diffuse axonal injury (DAI), is a concern as it is a severe and traumatic brain injury caused by forces of acceleration and deceleration. DAI is a common cause of unconsciousness, persistent vegetative state, and profound neurologic deficits following head trauma, particularly in high-speed car accidents. MRI is a valuable tool in diagnosing DAI, as it can detect lesions in the white matter tracts and is more sensitive than CT scans. However, the MRI findings in DAI

Figure 2: Generated responses to the example prompt in Figure 1.

Model	**Decoding time (s)**
llama-70b (4-bit)	37.8
llama-8b	4.8
pisco-llama (8B)	3.1
mistral-7b	4.9
pisco-mistral (7B)	3.1

Table 3: Average decoding time per model when using 2 GPUs (llama-70b) or 1 GPU (otherwise).

3.2 Generation times

The average response time per question per model is shown in Table 3. Although these times could be further optimized by tuning the batch size, for example, it is clear that the quantized llama-70b model is substantially slower than llama-8b, for instance, while it takes up more GPU memory. In addition, the PISCO models that operated on compressed inputs actually spent less time on decoding than those that did not. Since context compression can be applied to all abstracts in advance, it does not add to the total generation time.

3.3 Official evaluation scores

In this part, we discuss the official evaluation results for our submitted runs (Gupta et al., 2025). The total number of evaluated answers is 30 for our first two runs, 28 for the third and fourth runs, and 29 for the fifth run. For details on the evaluation setup using a combination of

Run	**Accuracy**	**Precision**	**Redundancy**	**Harmfulness**
llama-70b (4-bit)	100	71.55	14.12	0.67
llama-8b	100	60.35	19.80	3.10
pisco-llama (8B)	92.86	72.20	9.52	0
mistral-7b	92.86	63.78	15.70	0.30
pisco-mistral (7B)	82.76	69.31	8.05	2.53

Table 4: Answer quality results.

Embedding model	**Run**	**Recall (S+R)**	**Recall (R)**	**Recall (R+B)**
SentenceTransformer	llama-70b (4-bit)	28.00	33.33	34.67
	llama-8b	29.00	37.00	43.00
	pisco-llama (8B)	2.50	19.64	21.79
	mistral-7b	28.57	31.07	36.43
	pisco-mistral (7B)	11.38	17.59	18.28
SimCSE	llama-70b (4-bit)	29.00	34.67	38.33
	llama-8b	31.00	37.67	42.67
	pisco-llama (8B)	2.50	20.00	22.50
	mistral-7b	27.86	31.07	38.93
	pisco-mistral (7B)	10.69	18.97	19.31

Table 5: Answer completeness results (S=Supported, R=Required, B=Borderline).

human and automated judgments, see Gupta et al. (2026).

Answer quality Table 4 shows the percentage of acceptable answers generated by each system, as well as the fraction of required, unnecessary, and harmful sentences in the given answers. We observe that compared to llama-7b, the quantized llama-70b model generated more required, less unnecessary, and less harmful sentences. This is in line with our initial analysis and the notion that bigger models tend to generate better answers. Furthermore, all answers given by both systems were deemed acceptable, whereas the PISCO models generated a few less acceptable answers than their uncompressed counterparts. Even so, the answers generated by the PISCO models contained more required and less redundant sentences in proportion to the total number of generated sentences, which was lower for these systems.

Answer completeness Answer recall measures how many of the answer aspects (pooled across all submitted runs from all participating teams) are covered in one answer to a question. Table 5 shows the answer completeness results for two sentence clustering methods in three settings: the sentences that were considered for clustering had to be jugded required and supported by evidence (S+R), required (R), or at least relevant (R+B). Because pisco-llama rarely provided citations, its recall in the first setting is very low. In general, both PISCO models achieved low scores due to the limited number of sentences and, consequently, answer aspects in their outputs. The highest recall was achieved by llama-8b, which generated the longest responses.

Citation quality Coverage is defined as the percentage of relevant sentences that are supported by appropriate citations. Table 6 shows the coverage per system, as well as the probability that any citation generated by that system supports or contradicts the corresponding statement (according to the human judgments). Based on the scores in this table, we conclude that pisco-llama is particularly ill-suited to the task of reference attribution. Not only did the system generate few citations; these citations also rarely provided evidence to support the corresponding claim. Unfortunately, the majority of citations generated by pisco-mistral did not support the corresponding statement

Run	Coverage	Support Rate	Contradict Rate
llama-70b (4-bit)	87.62	79.03	2.04
llama-8b	73.83	92.38	3.49
pisco-llama (8B)	9.52	13.39	0.89
mistral-7b	72.75	83.13	5.23
pisco-mistral (7B)	45.8	41.68	4.30

Table 6: Citation quality results.

Run	Recall	Precision
llama-70b (4-bit)	5.65	97.36
llama-8b	5.35	93.43
pisco-llama (8B)	0.48	16.07
mistral-7b	4.09	92.86
pisco-mistral (7B)	1.88	60.63

Table 7: Document relevance results.

either. In contrast, mistral-7b achieved a substantially higher support rate of 83%. llama-8b performed even better in the sense that it had the highest support rate, namely 92%, while it generated fewer contradicting citations per statement. Both models covered about 3 in 4 relevant sentences with citations. The quantized llama-70b model covered more relevant sentences with citations, but this came at the cost of a lower support rate.

Document relevance Table 7 displays the average recall and precision of the documents cited per answer. These scores are calculated against all documents judged relevant based on pooling across all submitted runs from all participating teams. This results in a large number of relevant documents, which explains why the average recall scores all seem relatively low. These results further demonstrate that the documents cited by the PISCO-based systems, and pisco-llama in particular, are less relevant than those cited by the standard RAG systems. In addition, we observe that llama-8b cited more distinct relevant documents per answer than mistral-7b, with a similar high precision. The system that used a quantized llama-70b model as generator achieved the highest recall and a very high precision of 97%. These scores also indicate that the retrieval setup, which was the same across all runs, was effective despite the lack of domain-specific training.

4 Conclusions and discussion

In conclusion, we evaluated the performance of five LLMs in task B of the TREC 2025 BioGen track, which is about long-form biomedical question answering with reference attribution. Here, we list our main findings. First, we found the zero-shot generalization of the PISCO models to be poor, mainly because these models were trained to generate very short outputs. Second, expectedly, we found that model size and architecture influence response quality. In particular, we found that a quantized llama-70b model could generate more reliable answers than a smaller llama-8b model due to its enhanced instruction-following capacities. Third, we found that models that generated higher-quality responses spent more time on the generation of these responses, affecting the time a user has to wait. These findings highlight the need for a more sophisticated approach to combining reference attribution with context compression — in other words, combining LLM reliability with efficiency. Therefore, in future work, we aim to investigate the dedicated training of larger models that make use of context compression rather than the zero-shot model application of smaller ones.

Acknowledgments

Jan Bakker and Jaap Kamps are supported by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan

Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2026. Bioace: An automated framework for biomedical answer and citation evaluations. *Preprint*, arXiv:2602.04982.

Deepak Gupta, Dina Demner-Fushman, William R. Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of TREC 2024 biomedical generative retrieval (BioGen) track. In *Proceedings of the Thirty-Third Text REtrieval Conference, TREC 2024, Gaithersburg, MD, USA, November 18-22, 2024*, volume 1329 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Deepak Gupta, Dina Demner-Fushman, William R. Hersh, Steven Bedrick, and Kirk Roberts. 2025. Overview of TREC 2025 biomedical generative retrieval (BioGen) track. In *Proceedings of the Thirty-Fourth Text REtrieval Conference, TREC 2025, Online, December 11-12, 2025*, NIST Special Publication. National Institute of Standards and Technology (NIST).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Maxime Louis, Hervé Déjean, and Stéphane Clinchant. 2025. PISCO: Pretty simple compression for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15506–15521, Vienna, Austria. Association for Computational Linguistics.

Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Ragnarök: A reusable rag framework and baselines for trec 2024 retrieval-augmented generation track. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I*, page 132–148, Berlin, Heidelberg. Springer-Verlag.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, Miami, Florida, USA. Association for Computational Linguistics.

David Rau, Shuai Wang, Hervé Déjean, Stéphane Clinchant, and Jaap Kamps. 2025. Context embeddings for efficient answer generation in retrieval-augmented generation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 493–502, New York, NY, USA. Association for Computing Machinery.