TREC BioGen 2025: A Retrieval and NLI-Based Approach for Biomedical Evidence Grounding

Riccardo Lunardi
riccardo.lunardi@uniud.it
University of Udine
Udine, Italy

Riccardo Zamolo
zamolo.riccardo@spes.uniud.it
University of Udine
Udine, Italy

Maria Elena Zuliani
zuliani.mariaelena@spes.uniud.it
University of Udine
Udine, Italy

Stefano Mizzaro
stefano.mizzaro@uniud.it
University of Udine
Udine, Italy

Vincenzo Della Mea
vincenzo.dellamea@uniud.it
University of Udine
Udine, Italy

Kevin Roitero
kevin.roitero@uniud.it
University of Udine
Udine, Italy

Abstract

This technical report presents the system developed for the TREC 2025 Biomedical Generative Retrieval Track. Our approach integrates both sparse and dense retrieval, LLM-based reranking, and Natural Language Inference (NLI) to identify both supporting and contradicting evidence for grounded answer generation, evaluating the interplay between them. We submitted 14 runs, four for Task A and ten for Task B, aimed to analyze how different retrieval and grounding configurations impact the factuality and reliability of biomedical answer generation.

Keywords

TREC, BioGen, LLMs, Information Retrieval

1 Introduction

The TREC BioGen Track was established to address the reliability of Large Language Models (LLMs) in the biomedical domain. Following the pilot task organized in 2024 [3], the 2025 edition introduces a reference attribution task, aimed at mitigating the generation of false statements when answering biomedical questions. Specifically, a grounding generation task was added to enhance the trustworthiness of LLM-generated answers by requiring citations from reliable sources.

Our goal for this track is to develop a pipeline capable of retrieving relevant biomedical documents and generating accurate, well-supported answers to complex biomedical questions. Our system leverages a combination of traditional Information Retrieval (IR) techniques, including sparse and dense retrieval, and advanced reranking methods. Furthermore, we integrate Natural Language Inference (NLI) models and LLMs to verify the support of retrieved documents and ensure the quality of the generated responses.

2 Methods

In this section, we describe the methods and techniques employed in our system. Our approach consists of several key components that are used for both tasks, specifically document retrieval and reranking.

2.1 Retrieval

We index the document collection processed by TREC using PySerini [4], creating both sparse and dense indices. The sparse index is built using BM25 [7], while the dense index relies on the `castorini/tct_colbert-v2-hnp-msmarco` model to generate dense vector representations. For both indices, we create two versions: one indexing full documents and another indexing passages. Full documents are indexed as provided, for a total of 26,805,982 entries. To build the passage-level index, each document is segmented into coherent text units based on sentence boundaries. Each passage contains approximately 500 characters, with an overlap of 50 characters to preserve context continuity. This segmentation procedure results in 94,691,711 passages.

2.2 Reranking

Recent studies [1] have shown that LLM-based rerankers can substantially improve retrieval quality. For this reason, we employ ZephyrReranker [6], which uses the `castorini/rank_zephyr_7b_v1_full` model. We choose ZephyrReranker due to its strong out-of-the-box performance and its ability to operate effectively without task-specific fine-tuning, which is particularly advantageous in an evaluation setting such as TREC BioGen. In our pipeline, the reranker receives the top-*N* documents or passages from the initial retrieval stage and reorders them based on their estimated relevance to the query.

2.3 Task A

Task A consists of retrieving documents that ground each sentence of the answer with appropriate PubMed documents. The system is also expected to provide PMIDs that contradict the statements made in each answer sentence.

We approach Task A by first expanding the queries using microsoft/BioGPT-Large, a biomedical language model. This expansion helps to capture the nuances of biomedical terminology and improve retrieval effectiveness.

The expanded query is then used to retrieve the top-50 relevant documents. For each document, we use the facebook/bart-large-mnli NLI classification model to determine whether the document supports, contradicts, or is neutral with respect to the main query. Finally, for each topic, we select up to 3 documents that support the query and 3 documents that contradict it, ensuring a balanced representation of evidence. If fewer than three supporting or contradicting documents are available, we include all available ones. Figure 1 shows the system adopted for Task A.

2.4 Task B

Task B involves generating a grounded answer to a biomedical question, for which we adopt a passage-level approach. We retrieve the top-100 passages, using either sparse or dense retrieval. This cutoff was determined through preliminary internal experiments, which indicated that 100 passages offer a good balance between evidence diversity and compatibility with the context limits of the generative models.

After retrieval, the passages are reranked using ZephyrReranker. We then use the NLI model to classify each passage and select the top-10 supporting and top-10 contradicting pieces of evidence. This curated collection of aligned and conflicting evidence forms the input context for the generative stage.

For the generation step, we evaluate three models: `gpt-4o-mini` [5], `Llama-3.3-70B-Instruct`, and `Llama-3.1-8B-Instruct` [2]. The selected passages are concatenated into a single context block and provided to the LLM with a structured system prompt. The prompt instructs the model to produce a concise answer of maximum 250 words, explicitly acknowledge both supporting and contradicting evidence, and include 1–3 PMIDs at the end of every sentence. The exact prompt used in our experiments is shown below:

```
You are a medical assistant expert. Your task is to write a
concise, factually grounded answer based on the information
below.
Claim: [question]
Supporting evidence: [supporting_evidence]
Contradicting evidence: [contradicting_evidence]
Instructions:
      • Use accurate and medical tone to write an answer
        based on the evidence provided. You can merge two
        similar sentences in one and add related PMIDs
      • Each sentence must include citations from 1 to 3
        document PMIDs, written in square brackets (e.g.,
        [30311221, 21306691]) at the end of the sentence
      • Each sentence must end with citations list followed
        by a dot (.)
      • Do not speculate or include any information that is
        not explicitly supported by the evidence documents
      • The answer must be no longer than 250 words
Answer:
```

Finally, the raw output generated by the LLM is parsed to strictly adhere to the submission format required by the TREC BioGen Track. The full pipeline for Task B is illustrated in Figure 2.

2.5 Submissions

We submitted a total of 14 runs to the TREC 2025 BioGen Track: four runs for Task A and ten runs for Task B. A detailed overview of the submitted configurations is provided in Table 1.

For Task A, our submissions aim to evaluate the impact of the retrieval method (Sparse vs. Dense) and the effectiveness of the reranking stage.

Table 1: Details of the submitted runs. The table specifies the retrieval method (Sparse/Dense), the usage of the Zephyr reranker, and the generative model used for Task B.

Task	Run ID	Retrieval	Rerank	Generator Model
A	1	Sparse	-	*N/A (Retrieval Only)*
	2	Sparse	✓	
	3	Dense	-	
	4	Dense	✓	
B	1	Sparse	-	Llama-3.1-8B-Instruct
	2	Sparse	✓	
	3	Dense	-	
	4	Dense	✓	
	5	Sparse	-	Llama-3.3-70B-Instruct
	6	Sparse	✓	
	7	Dense	-	
	8	Dense	✓	
	9	Sparse	-	gpt-4o-mini
	10	Dense	-	

For Task B, we expanded the analysis to include the generation component. We maintained the same retrieval and reranking variations used in Task A, but introduced different LLMs to generate the final grounded answers.

3 Results

This section presents the experimental results for both tasks. We first analyze Task A, which evaluates evidence retrieval performance, and then report the end-to-end generation and citation results for Task B.

3.1 Task A

Table 2 summarizes the system performance on Task A. The primary finding concerns the trade-off between retrieval methods: sparse retrieval (Run 1) outperforms dense retrieval (Run 3) in identifying supporting evidence, achieving a precision of 0.41 compared to 0.27. In contrast, dense retrieval is slightly more effective at surfacing contradicting evidence, yielding higher Contradict Precision (0.03) and Recall (0.09) than the sparse baseline, although absolute values remain low.

The integration of the Zephyr reranker has limited impact on the system. For sparse retrieval (Run 2), it slightly increases Support Recall but degrades other metrics. For dense retrieval (Run 4), performance decreases substantially across all metrics, with only marginal changes in contradiction scores.

Finally, a comparison of automatic and human evaluation for Run 1 shows that automatic metrics closely match the "relaxed" human setting. However, under "strict" evaluation, Support Precision and Recall drop to 0.15 and 0.17, respectively, indicating that automatic matching may overestimate performance when stricter relevance criteria are applied.

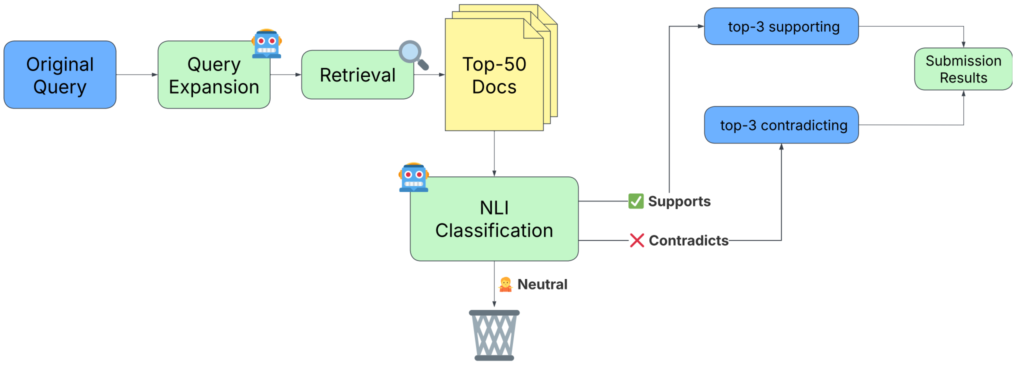

Figure 1: Overview of the system pipeline for Task A.

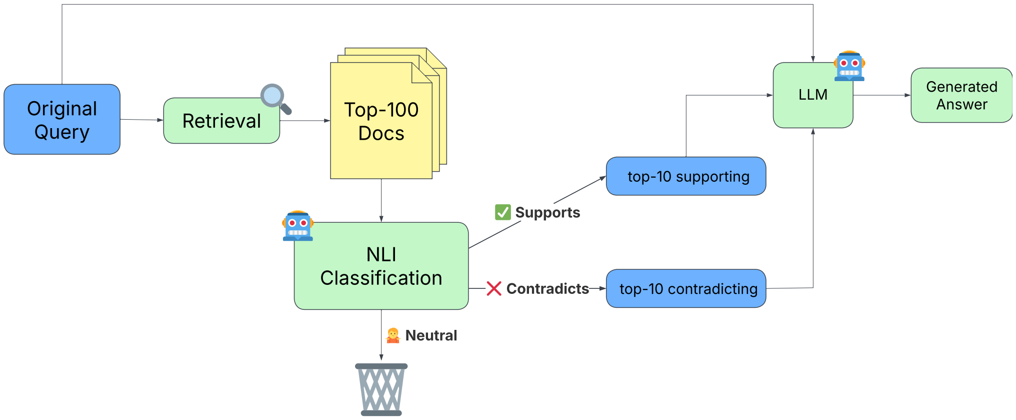

Figure 2: Overview of the system pipeline for Task B.

Table 2: Task A results evaluating retrieval, reranking, and evaluation methods.

Run	Retrieval	Rerank	Evaluation	Support		Contradict	
				Precision	Recall	Precision	Recall
1	Sparse	-	Automatic	0.41	0.42	0.01	0.03
1	Sparse	-	Human (Relaxed)	0.40	0.43	0.03	0.03
1	Sparse	-	Human (Strict)	0.15	0.17	0.03	0.03
2	Sparse	✓	Automatic	0.39	0.45	0.02	0.02
3	Dense	-	Automatic	0.27	0.19	0.03	0.09
4	Dense	✓	Automatic	0.17	0.09	0.03	0.07

3.2 Task B

The performance metrics for Task B are shown in Table 3. Since trends are consistent across generative models, we focus here on the effect of retrieval and reranking using the automatic metrics available for all runs.

We observe a similar trend to Task A: sparse retrieval outperforms dense retrieval in Precision (0.91 vs. 0.84), Recall (0.36 vs. 0.33), and Citation Support Rate (0.77 vs. 0.48). While overall contradiction rates remain low, dense retrieval is more effective at surfacing contradictory evidence, producing a Citation Contradict Rate of 0.06 compared to the sparse baseline's 0.01. Furthermore, the integration of the Zephyr reranker generally degrades the automatic Precision and Recall scores, only slightly increasing Citation Coverage.

Table 3: Task B results categorized by retrieval configuration and generative model.

Retrieval / Model	Rerank	Precision	Recall	Citation Coverage	Citation Support Rate	Citation Contradict Rate
			Retrieval Performance			
Dense	-	0.84	0.33	0.49	0.48	0.07
Dense	✓	0.81	0.32	0.57	0.46	0.06
Sparse	-	0.91	0.36	0.80	0.77	0.01
Sparse	✓	0.89	0.35	0.83	0.74	0.02
			Generative Model Performance			
Llama-3.1-8B-Instruct		0.87	0.36	0.63	0.60	0.02
Llama-3.3-70B-Instruct		0.84	0.31	0.77	0.60	0.05
gpt-4o-mini		0.90	0.36	0.50	0.68	0.03

Evaluating the generative models individually, `gpt-4o-mini` achieves the highest overall Precision, Recall, and Citation Support Rate, though it exhibits the lowest Citation Coverage. In contrast, `Llama-3.3-70B-Instruct` produces lower Precision and Recall but excels in citing evidence, reaching the highest Citation Coverage and Citation Contradict Rate. The smaller `Llama-3.1-8B-Instruct` model offers a competitive middle ground, maintaining strong Precision and Recall while balancing the citation metrics.

4 Conclusions

In this technical report, we present the University of Udine team's approach for the TREC 2025 BioGen Track. Our pipeline combines sparse and dense retrieval with NLI-based filtering to ground LLM generations in reliable biomedical literature.

Based on our implemented configurations, we observe a clear advantage for sparse retrieval, which consistently outperforms our dense retrieval setup in identifying accurate supporting evidence. Furthermore, the Zephyr reranker, as deployed in our current pipeline, proved ineffective and generally degraded the retrieval metrics.

Finally, operating downstream of these specific retrieval settings, the evaluated generative models exhibit distinct behaviors: `gpt-4o-mini` favors higher overall precision and factual support, whereas `Llama-3.3-70B-Instruct` prioritizes broader citation coverage. These results suggest that, in our setup, retrieval configuration has a stronger impact than model choice, and future improvements should focus primarily on retrieval and reranking components.

Acknowledgments

This work was supported in part by the ESF+ 2021/2027 Regional Program of the Autonomous Region Friuli Venezia Giulia (PPO 2023, Program No. 22/23 - LINE A: PhD programmes) and the Region FVG Project "Supporting the diagnosis of rare diseases (MR) through artificial intelligence" (2023–2026, Project A: F53C22001770002)

References

[1] Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025. How Good are LLM-based Rerankers? An Empirical Analysis of State-of-the-Art Reranking Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 5693–5709. doi:10.18653/v1/2025.findings-emnlp.305

[2] Aaron Grattafiori, Abhimanyu Dubey ..., and Zhiyu Ma. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[3] Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of TREC 2024 Biomedical Generative Retrieval (BioGen) Track. arXiv:2411.18069 [cs.IR] https://arxiv.org/abs/2411.18069

[4] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2356–2362. doi:10.1145/3404835.3463238

[5] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[6] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! arXiv:2312.02724 [cs.IR] https://arxiv.org/abs/2312.02724

[7] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389. doi:10.1561/1500000019