RAGonauts at SemEval-2026 Task 8: BM25 Retrieval with Last-Turn Query Formulation for Multi-Turn RAG Conversations

Rajalakshmi Sivanaiah, Angel Deborah S, Karthik Raja C, Rithika S
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai 603110, Tamil Nadu, India
rajalakshmis@ssn.edu.in angeldeborahs@ssn.edu.in
karthikraja2210450@ssn.edu.in rithika2210979@ssn.edu.in

Abstract

This paper describes the submission to Task A of SemEval-2026 Task 8: MTRAGEval which evaluates passage retrieval for multi-turn Retrieval-Augmented Generation (RAG) conversations across multiple knowledge domains. The task requires retrieving relevant supporting passages given conversational history, where user queries often contain implicit references and incomplete contextual information. This paper proposes a lightweight and training-free retrieval framework based on BM25 ranking combined with conversational query formulation. Queries are derived from dialogue turns and retrieval is performed using domain-specific indices to preserve corpus relevance. Without neural retrievers or fine-tuning, our system achieves an nDCG@5 score of *0.2836* on the official evaluation set, ranking *33rd* on the leaderboard. This result demonstrates that sparse lexical retrieval remains an efficient and reproducible baseline for conversational RAG systems.

1 Introduction

Retrieval-Augmented Generation (RAG) systems enhance the factual reliability of language models by retrieving supporting evidence from external knowledge sources prior to response generation (Lewis et al., 2020). While significant progress has been achieved in single-turn retrieval, real-world information-seeking interactions frequently occur as multi-turn conversations. In conversational settings, follow-up queries often depend on previous dialogue context through co-reference, ellipsis, or topic continuation, making passage retrieval substantially more challenging than traditional ad-hoc search.

SemEval-2026 Task 8: MTRAGEval addresses this challenge by evaluating retrieval performance in multi-turn RAG conversations across multiple knowledge domains, building on the MTRAG benchmark originally introduced by Katsis et al. (2025) (Katsis et al., 2025). Task A focuses specifically on conversational passage retrieval, where systems must return a ranked list of ten relevant passages given the conversation history and a domain-specific corpus. The benchmark includes four diverse collections: Wikipedia (CLAPNQ), technical documentation (CLOUD), financial question answering (FIQA), and government documents (GOVT).

This paper's objective is to establish a strong and computationally efficient lexical retrieval baseline using a lightweight architecture. Retrieval is performed using independent domain-specific indices, ensuring consistent corpus statistics and preventing cross-domain interference. The proposed system is entirely training-free, requires no GPU resources, and relies solely on the datasets provided by the task organizers.

Despite its simplicity, the proposed method obtains an nDCG@5 score of **0.2836** on the evaluation data. The observations show the significant variability in performance across the domains, which points to the challenges posed by the implicit conversational context and vocabulary. These results show the efficiency, interpretability, and reproducibility of the proposed sparse methods for the multi-turn conversational RAG model.

2 Background

2.1 Task Description

Task A of SemEval-2026 Task 8: MTRAGEval addresses passage retrieval in multi-turn Retrieval-Augmented Generation (RAG) conversations. Each task provides a dialogue history consisting of alternating user queries and system responses together with a target domain collection. Given the conversation context up to the current user turn, systems must retrieve a ranked list of ten relevant passages from the corresponding corpus.

A central challenge arises from conversational

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1938–1943

dependencies between turns, where follow-up questions often omit entities introduced earlier in the dialogue. As a result, retrieval systems must interpret implicit contextual information rather than rely solely on explicit keyword matching. The evaluation set additionally introduces an *underspecified* answerability class (Katsis et al., 2025), containing queries whose intent cannot be uniquely determined from the available context.

2.2 Dataset

The task is based on the MTRAG-UN benchmark (Katsis et al., 2025), which contains human-authored multi-turn conversations across four English-language domains. Table 1 summarises corpus statistics and task distribution.

Documents are segmented into fixed-length passages with overlap to preserve local context. Conversations span multiple turns and include factual queries, clarifications, and topic continuation, making retrieval sensitive to evolving conversational context.

Domain	Passages	Train Tasks	Eval Tasks
ClapNQ (Wikipedia)	183,408	208	142
Cloud (Tech Docs)	61,022	217	131
FiQA (Finance)	49,607	216	77
Govt (Government)	72,422	201	157
Total	366,459	**842**	**507**

Table 1: Corpus statistics and task distribution across domains in the MTRAG-UN benchmark.

2.3 Evaluation Metrics

Retrieval effectiveness is measured using Recall@k and normalised Discounted Cumulative Gain (nDCG@k) for $k \in \{1, 3, 5, 10\}$ based on human relevance judgements. Metrics are computed per conversational turn and aggregated across domains. The official leaderboard metric is **nDCG@5**, emphasising accurate ranking of highly relevant passages.

2.4 Related Work

Sparse Retrieval. Lexical methods such as BM25 (Robertson et al., 1994) remain strong retrieval baselines due to their efficiency and effectiveness when query and document vocabularies overlap.

Dense Retrieval. Neural retrievers, including Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) and embedding-based models such as BGE (Xiao et al., 2023), map queries and documents into semantic vector spaces, enabling matching beyond exact lexical similarity.

Conversational Retrieval. The work on conversational search, especially the TREC CAsT benchmark (Dalton et al., 2020), has shown the benefits of co-reference resolution and ellipsis resolution through query rewriting. The original MTRAG benchmark (Katsis et al., 2025) and the subsequent MTRAG-UN benchmark have also found significant benefits in rewriting conversational queries before the retrieval process.

3 System Description

Our retrieval pipeline consists of three stages: (i) conversational query construction, (ii) domain-aware BM25 retrieval, and (iii) result formatting for evaluation. The Figure 1 represents the proposed architecture of the system.

This design prioritises simplicity, interpretability, and reproducibility. By avoiding training and external resources, the system provides a controlled setting to analyse the impact of conversational query formulation on retrieval performance. Such lightweight baselines are particularly valuable in shared tasks, where they serve as reference points for understanding the gains achieved by more complex architectures.

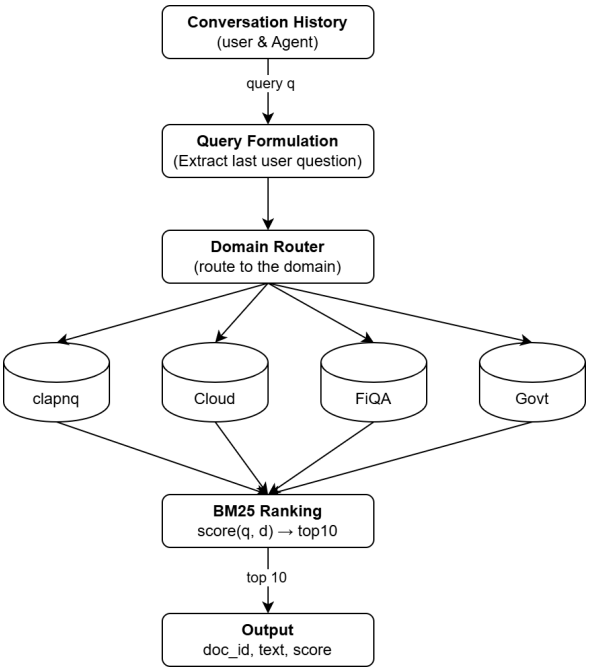

Figure 1: Proposed System Architecture

3.1 Query Formulation

Transforming conversational dialogue into an effective search query is central to multi-turn retrieval. We implement three query construction strategies:

1. **last_question**: uses only the most recent user utterance,
2. **all_questions**: concatenates all user turns,
3. **full_conversation**: includes both user and system turns.

The `last_question` strategy is used for the official submission, as the latest user turn typically represents the most specific information need while avoiding noise introduced by earlier dialogue content. All text is lowercased and tokenised using alphanumeric word boundaries.

3.2 BM25 Retrieval

We employ the BM25Okapi implementation from the `rank-bm25` library (Brown, 2021). Separate BM25 indices are constructed for each domain corpus, selected dynamically using the task collection identifier. This design prevents cross-domain interference and preserves domain-specific term statistics used during scoring.

Document Representation. For each document d, the title and passage text are concatenated prior to indexing is in Equation 1:

$$\text{tokens}(d) = \text{tok}(\text{lower}(t_d \oplus p_d)), \qquad (1)$$

where $\oplus$ denotes concatenation and tokenisation splits text using the regular expression `\b\w+\b`.

Ranking Function. Document relevance is computed using the standard BM25 scoring function in the Equation 2:

$$\text{BM25}(q,d) = \sum_{w \in q} \text{IDF}(w) \times \frac{f(w,d)(k_1+1)}{f(w,d) + k_1 \left(1 - b + b\frac{|d|}{|d_{\text{avg}}|}\right)} \qquad (2)$$

with default parameters $k_1 = 1.5$ and $b = 0.75$. The top ten passages ranked by score are returned as retrieval contexts.

4 Experimental Setup

4.1 Data

We use only the datasets provided by the task organisers, without external resources. The training split (`human/generation_tasks/RAG.jsonl`) contains 842 conversational retrieval tasks with relevance judgements (qrels) used for local evaluation. The official evaluation set (`rag_taskAC.jsonl`) consists of 507 tasks with labels withheld during submission. Table 1 shows the domain-wise distribution.

4.2 Implementation

The system is implemented in Python 3.10 using the `rank-bm25` library (Brown, 2021). Separate BM25 indices are constructed for each domain corpus. Documents are indexed by concatenating titles and passage text, followed by lowercasing and tokenisation using word boundaries. We do not apply stop-word removal, stemming, or lemmatization, in order to preserve domain-specific terminology and maintain a fully training-free and reproducible baseline. Tokenisation is performed using the regular expression `\b\w+\b`, which retains alphanumeric tokens while removing punctuation. Default BM25 parameters ($k_1 = 1.5$, $b = 0.75$) are used without tuning.

All experiments are conducted on a CPU-only machine. Index construction ranges from approximately 2–9 minutes depending on corpus size, and inference on the evaluation set completes in about 44 minutes. Submission outputs are validated using the official `format_checker.py`. [1]

4.3 Query Strategies and Evaluation

We evaluate three conversational query strategies: (i) `last_question`, (ii) `all_questions`, and (iii) `full_conversation`. Based on training-set experiments, the `last_question` strategy is used for the official submission.

Performance on the training split is measured using Recall@k and nDCG@k ($k \in \{1, 3, 5, 10\}$) following the official evaluation protocol. Scores on the evaluation set correspond to the results reported by the task organisers.

[1] Code is publicly available at https://github.com/itskarthik17/sem-eval-task8.

5 Results

5.1 Official Evaluation Results

Table 2 reports the official Task A results on the blind evaluation set. Our BM25-based retrieval system achieves an nDCG@5 score of **0.2836**, ranking **33rd** out of all submissions on the official leaderboard. Although neural and hybrid approaches achieve higher absolute performance, this result demonstrates that a lightweight, training-free lexical retriever provides a reliable baseline for multi-turn conversational retrieval.

System	nDCG@5
Top Performing System	0.5776
RAGonauts (BM25)	**0.2836**

Table 2: Official Task A leaderboard results on the blind evaluation set.

The evaluation set additionally introduces an *underspecified* answerability class, increasing retrieval difficulty by including conversational turns where the intended information need cannot be uniquely determined from the available context. This particularly affects lexical retrieval methods such as BM25, which rely on explicit term matching and cannot infer missing entities or implicit references. As a result, underspecified queries often lead to partially relevant or topically related passages rather than precise matches, contributing to the observed performance gap compared to more advanced retrieval approaches.

5.2 Training Set Performance

Since relevance judgements are available only for the training split, we report detailed retrieval metrics on this data. Table 3 summarises Recall and nDCG scores averaged across all domains.

Metric	@1	@3	@5	@10
Recall	0.078	0.149	0.194	0.262
nDCG	0.172	0.158	0.175	0.204

Table 3: Training-set retrieval performance (macro-average over 842 tasks).

Our nDCG@10 score of 0.204 closely matches the published BM25 last-turn baseline reported in the MTRAG-UN benchmark (Katsis et al., 2025), indicating consistent behaviour despite differences in preprocessing and implementation choices.

5.3 Per-Domain Performance

Table 4 presents domain-wise retrieval results. Performance varies across domains, reflecting differences in document structure and vocabulary distribution.

Domain	Tasks	Recall@10	nDCG@10
Govt	201	0.339	0.267
ClapNQ	208	0.291	0.230
Cloud	217	0.264	0.213
FiQA	216	0.153	0.104
Overall	**842**	**0.262**	**0.204**

Table 4: Per-domain training-set results using the `last_question` strategy.

The GOVT domain achieves the strongest performance, likely due to consistent terminology and structured document language that aligns well with lexical matching. In contrast, the FIQA domain remains the most challenging, where informal financial discussions reduce lexical overlap between queries and relevant passages.

Prior work on the MTRAG-UN benchmark shows that dense retrieval and query rewriting techniques substantially improve conversational retrieval performance (Katsis et al., 2025). Our results therefore establish a lightweight sparse-retrieval baseline against which more computationally intensive approaches may be compared.

6 Analysis

6.1 Query Strategy Ablation

Table 5 presents an exploratory comparison of query formulation strategies on a subset of 50 training tasks. Given the relatively small sample size, the results should be interpreted as indicative rather than conclusive.

The `all_questions` strategy achieves slightly higher performance on this subset, suggesting that earlier conversational turns can provide useful contextual signals. However, we adopt the `last_question` strategy for the official submission as a minimal and robust baseline, as it avoids introducing potentially noisy or redundant context from earlier turns and reflects a common standard in conversational retrieval baselines.

Incorporating all user questions results in a modest performance improvement, indicating that earlier conversational turns provide useful contextual

Strategy	nDCG@10
last_question (submission)	0.204
all_questions	0.218
full_conversation	0.211

Table 5: Query formulation ablation on a 50-task training subset.

information for retrieval. However, including agent responses does not consistently improve performance, suggesting that additional dialogue content may introduce noise when responses contain acknowledgements or non-informative text. These findings highlight the trade-off between contextual coverage and query specificity in conversational retrieval.

While this analysis focuses on the CLAPNQ domain, performance differences across domains—particularly the lower results observed for FIQA—suggest that domain-specific conversational patterns and vocabulary mismatches may further impact retrieval effectiveness. Extending error analysis to additional domains such as FIQA remains an important direction for future work.

6.2 Error Analysis

To better understand retrieval failures, we manually examined 30 instances from the CLAPNQ domain as a representative sample for qualitative analysis. Three recurring error patterns were observed:

1. **Co-reference ambiguity (43%):** follow-up questions frequently refer to entities introduced in earlier turns, which lexical retrieval cannot resolve without explicit entity mentions.

2. **Vocabulary mismatch (30%):** semantic variations between queries and documents reduce token overlap, limiting BM25 effectiveness.

3. **Underspecified queries (27%):** queries lacking sufficient contextual detail lead to retrieval of topically related but irrelevant passages.

These observations indicate that conversational understanding and semantic matching remain key challenges for sparse retrieval methods.

Overall, the results highlight that while sparse retrieval methods remain competitive as baselines, they are fundamentally limited in handling implicit conversational context and semantic variation. These limitations become more pronounced in domains with informal language, such as FIQA, and in cases involving underspecified queries. This suggests that future improvements are likely to come from integrating conversational query rewriting and semantic retrieval techniques, particularly in multi-domain conversational settings.

7 Conclusion

This paper presented the RAGonauts submission to SemEval-2026 Task 8 Task A, introducing a BM25-based retrieval framework for multi-turn conversational RAG. Despite relying solely on lexical matching and requiring no training or external resources, the proposed system achieves competitive performance while operating entirely on CPU.

Our analysis shows that conversational dependencies and vocabulary mismatch remain primary limitations of sparse retrieval approaches. Future work will explore conversational query rewriting, dense retrieval models, and hybrid sparse–dense retrieval methods to better capture implicit conversational context and improve retrieval robustness.

References

Dorian Brown. 2021. rank-bm25: A collection of BM25 algorithms in python.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. In *Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019)*. NIST.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3929–3938.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Preprint*, arXiv:2501.03468.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–548.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 249–266.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST.

Thomas Wolf, Lysandre Debut, Victor Sanh, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2021. Few-shot generative conversational query rewriting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1933–1937.