TDSA_UIT at SemEval-2026 Task 8: Hybrid Dense–Sparse Retrieval for Multi-Turn RAG Conversations

Phuoc-Thinh Dao[1,2], Quang-Dat Ha[1,2], Ngoc-Son Tran[1,2], Huu-An Nguyen[1,2], Hoang-Nam Nguyen[1,2], Duc-Vu Nguyen[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{25210038, 25210008, 25210033, 25210001,25210022}@ms.uit.edu.vn vund@uit.edu.vn

Abstract

This paper presents an investigation into enhancing information retrieval performance within multi-turn conversational systems, conducted under the framework of SemEval-2026 Task 8. We focus on addressing information-deficient queries in multi-turn dialogues, where critical information is often implicit or replaced by coreferences. We propose a retrieval pipeline that integrates query contextualization techniques with a hybrid search strategy and subsequent reranking. Experimental results on the MTRAG benchmark demonstrate that an optimal hybrid search configuration (30:30:30) combined with a reranker achieves promising performance, yielding a recall@10 of 0.48 and an nDCG@10 of 0.42. Our findings indicate that the integration of query context enrichment and diverse retrieval modalities enables the system to maintain robust retrieval accuracy across extended conversational scenarios and multiple domains.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved revolutionary success in natural language understanding and generation. However, these models still face inherent limitations, such as hallucinations, outdated internal knowledge, and a lack of domain-specific expertise. To address these challenges, Retrieval-Augmented Generation (RAG) has emerged as a promising solution by integrating external knowledge sources into the LLM's generation process. RAG plays a crucial role in enhancing the accuracy, timeliness, and reliability of responses while significantly reducing costs compared to fine-tuning models on new data (Lewis et al., 2020; Huang and Huang, 2024).

While RAG systems have made significant strides in single-turn retrieval tasks, research focus is increasingly shifting toward multi-turn conversations—a more pragmatic yet complex scenario. In this setting, systems are required not only to address isolated queries but also to maintain coherence and accuracy throughout a sequence of interactions (Ye et al., 2024). Benchmarks such as MTRAG (Katsis et al., 2025) have been developed to evaluate a system's capability to overcome real-world multi-turn challenges, including proactive retrieval, long-form responses, and cross-domain data.

However, implementing RAG in multi-turn conversations continues to face several significant challenges. First is the contextual dependency, as user queries are often not standalone but contain information or concepts referenced from previous turns. Second, the need for proactive retrieval arises as the required information evolves continuously throughout the conversation, necessitating the system to update relevant passages at each step. Finally, answerability remains a critical hurdle; current models often struggle to refrain from answering when the necessary knowledge is absent from the documents, leading to high hallucination rates in later dialogue turns (Cheng et al., 2025; Rosenthal et al., 2025). These challenges demand synchronized improvements in both retrieval components and contextual reasoning capabilities during text generation.

2 Task description

MTRAG (Katsis et al., 2025) is the first fully human-generated multi-turn conversational benchmark designed for the comprehensive evaluation of RAG systems. It comprises 110 conversations (totaling 842 turns) spanning four distinct domains: Wikipedia (CLAP_{NQ}), Finance (FiQA), Government (Govt), and Cloud technical documentation (Cloud). A key highlight of MTRAG is its reflection of real-world properties, such as proactive retrieval, long-form responses, unanswerable questions, and non-standalone queries that exhibit heavy contextual dependency.

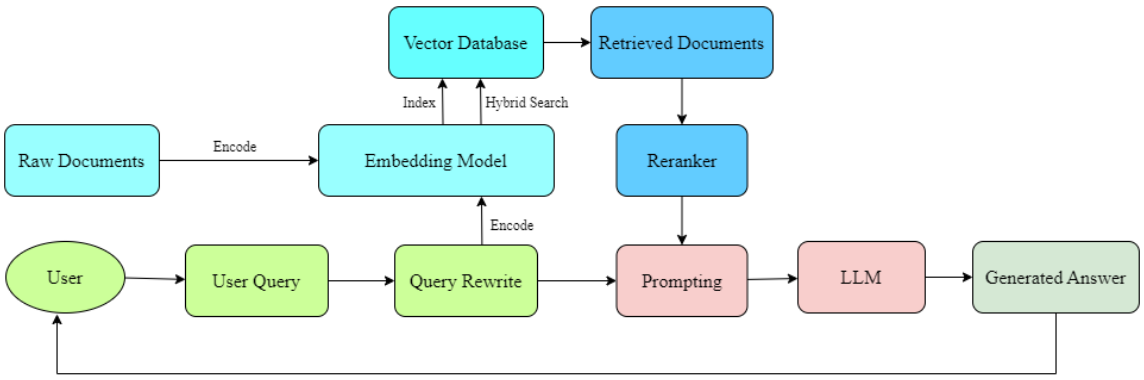

Figure 1: RAG pipeline

The MTRAGEval (Rosenthal et al., 2025) is built upon the MTRAG benchmark, focusing on three primary subtasks:

• Subtask A - Retrieval: Evaluates the system's ability to retrieve relevant passages for a query at the final turn of a multi-turn conversation.

• Subtask B - Generation with Reference Passages: Assesses the capability to generate responses grounded in a provided set of gold-standard reference passages.

• Subtask C - Full RAG: An end-to-end evaluation that integrates both passage retrieval and response generation steps.

The task requires models to process dialogue turns within a multi-turn conversation, encompassing the entire dialogue history alongside the final query. Two major challenges of this task include the ability to identify unanswerable questions to mitigate hallucinations, and the handling of later conversational turns, which are inherently complex due to their reliance on information or concepts from previous exchanges.

Recognizing that retrieval quality is a decisive factor and serves as the foundation for the entire RAG pipeline, this study focuses specifically on subtask A (Retrieval). Our objective is to enhance the accuracy of document retrieval in complex multi-turn conversations, thereby minimizing the propagation of misinformation to subsequent stages.

3 Experiment

In this study, we focus on enhancing the retrieval system for multi-turn conversations within the MTRAG benchmark. The overall system architecture is illustrated in Figure 1. The retrieval system is constructed based on the following components:

Query Rewriting Technique: For each dialogue turn, the current query is concatenated with the preceding conversation history and processed by a Large Language Model to generate a standalone retrieval query. This step facilitates context disambiguation and significantly improves retrieval effectiveness in multi-turn scenarios.

Embedding Model: We employ the *sentence-transformers/all-MiniLM-L6-v2* model, deployed locally, to transform both documents and queries into semantic vector representations within a continuous vector space.

Vector Database: Qdrant is utilized to store embeddings and perform dense retrieval based on the cosine similarity between query vectors and document vectors.

Reranker Model: We employ the *cross-encoder/ms-marco-MiniLM-L-6-v2* model with a parameter of $k = 10$ to re-rank the retrieved documents, thereby optimizing the priority order of the final results list.

We conduct a comparison between two retrieval strategies: *Dense search* and *hybrid search*.

Dense search is performed based on the cosine similarity between the query embeddings and document embeddings.

Hybrid search combines dense search with BM25-based sparse search. The configurations for the number of retrieved documents in hybrid search are denoted as $(k_{\text{dense}} : k_{\text{BM25}} : k_{\text{RRF}})$ and consist of two ratio groups:

- 1:1:1 Ratio: (10:10:10), (20:20:20), (30:30:30)
- 2:2:1 Ratio: (20:20:10), (40:40:20), (60:60:30)

To merge the results from the two retrieval

Table 1: Comparison of retrieval performance between dense search and hybrid search across the entire dataset

Search	Configuration	Recall				nDCG			
($k_{dense} : k_{BM25} : k_{RRF}$)		@1	@3	@5	@10	@1	@3	@5	@10
Dense	20	0.15	0.27	0.32	**0.39**	**0.37**	0.30	0.32	0.35
Hybrid	10:10:10	0.17	0.30	0.35	**0.38**	**0.39**	0.33	0.34	0.35
	20:20:20	0.17	0.30	0.35	**0.38**	**0.39**	0.33	0.34	0.35
	30:30:30	0.17	0.32	0.40	**0.48**	0.40	0.35	0.38	**0.42**
	20:20:10	0.17	0.30	0.36	**0.38**	**0.38**	0.33	0.35	0.36
	40:40:20	0.17	0.32	0.40	**0.47**	0.40	0.35	0.38	**0.41**
	60:60:30	0.17	0.32	0.40	**0.47**	0.40	0.35	0.38	**0.41**

Table 2: Detailed retrieval performance broken down by domain

Search	Domain	Recall				nDCG			
		@1	@3	@5	@10	@1	@3	@5	@10
Dense	CLAP_{NQ}	0.17	0.32	0.41	**0.50**	**0.47**	0.37	0.40	0.44
	Govt	0.15	0.28	0.32	**0.38**	**0.35**	0.30	0.32	0.34
	Cloud	0.15	0.26	0.29	**0.33**	**0.33**	0.28	0.28	0.30
	FiQA	0.13	0.22	0.27	**0.33**	**0.31**	0.26	0.27	0.20
Hybrid (30:30:30)	CLAP_{NQ}	0.19	0.37	0.45	**0.57**	0.50	0.42	0.44	**0.50**
	Govt	0.20	0.38	0.48	**0.55**	0.43	0.40	0.44	**0.47**
	Cloud	0.16	0.29	0.35	**0.42**	0.34	0.31	0.33	**0.36**
	FiQA	0.13	0.25	0.32	**0.37**	0.31	0.27	0.30	**0.32**

modalities, we apply the *Reciprocal Rank Fusion (RRF)* technique. The RRF score for a document d is calculated as follows:

$$\text{RRF}(d) = \sum_{i=1}^{n} \frac{1}{k + \text{rank}_i(d)}$$

where $\text{rank}_i(d)$ denotes the rank of document d within the i-th retrieval list, and the constant k is set to 60. This technique effectively leverages the strengths of both dense and sparse retrieval, generating a high-quality candidate list prior to the final ranking by the reranker model.

The experimental process was conducted on a total of 777 tasks across four domains in the MTRAG benchmark, including CLAP_{NQ} (208 tasks), FiQA (180 tasks), Govt (201 tasks), and Cloud (188 tasks). Since the data was pre-chunked, we proceeded directly with embedding and storage in the vector database.

The system's performance is evaluated using two primary metrics: *recall@k* and *nDCG@k*. To ensure an objective multi-domain assessment, these metrics are calculated as a weighted average across all 777 tasks rather than a simple arithmetic mean of the individual domain scores.

4 Results

Section 4 presents the experimental results of our system on the MTRAG benchmark across three primary evaluation scenarios:

First, we conduct a performance comparison of retrieval strategies. We compare the retrieval capabilities of the traditional method (*dense search*) against our enhanced approach (*hybrid search*). Various configurations with different numbers of retrieved documents were established to identify the model's saturation point (Table 1). Accuracy is measured at multiple cut-offs (@1, @3, @5, @10) to evaluate ranking effectiveness.

Second, we evaluate cross-domain stability (Table 2). The system is tested across four distinct domains in the MTRAG dataset to determine the model's adaptability to specialized linguistic characteristics in finance, government, and technology.

Third, we perform an ablation study on the reranker component (Table 3). To clarify the contri-

Table 3: Evaluation of the reranker model's impact on the retrieval performance of the hybrid search (30:30:30)

Reranker	Recall				nDCG			
	@1	@3	@5	@10	@1	@3	@5	@10
Yes	0.17	0.32	0.40	**0.48**	0.40	0.35	0.38	**0.42**
No	0.10	0.21	0.29	**0.38**	0.24	0.23	0.26	**0.30**

bution of the re-ranking step, we compare the performance of the hybrid search strategy (30:30:30) in two scenarios: with and without the use of a reranker model.

5 Discussion

The results in Table 1 demonstrate a clear superiority of the hybrid search strategy over traditional dense search. While dense search only achieves a recall@10 of 0.39, the integration of BM25 combined with the RRF fusion algorithm significantly boosts this metric, peaking at the 30:30:30 configuration with a recall@10 of 0.48 (an approximate 23% increase).

This proves that in multi-turn conversations, queries often contain both specific keywords and implicit semantic meanings. Relying solely on embeddings sometimes fails to capture critical entities, a deficiency that hybrid search effectively mitigates. However, when increasing the number of retrieved documents to the 60:60:30 mark, performance shows signs of saturation and slightly declines, suggesting that retrieving an excessive number of documents may introduce noise into the final ranking process.

Experiments across different domains in Table 2 indicate that the system performs most stably on the CLAP_{NQ} and Govt datasets. Notably, in the Govt set, hybrid search improves recall@10 from 0.38 to 0.55. Nevertheless, results for the FiQA (finance) domain remain relatively low, with a recall@10 of 0.37 and an nDCG@10 of only 0.32.

Compared to the results published in the original MTRAG paper, our findings exhibit a consistent trend: domains with highly technical terminology and complex reasoning, such as Finance (FiQA), remain the most significant challenges for general-purpose embedding models. This suggests a promising research direction in utilizing domain-specific embedding models.

The ablation study in Table 3 confirms that the reranker is an indispensable component for optimizing document priority. With the inclusion of the reranker, the nDCG@10 metric increases sharply from 0.30 to 0.42. This is of critical importance for practical RAG systems: surfacing accurate information at the top of the results helps the LLM mitigate hallucinations and conserve context window length, thereby enhancing the quality of the final response.

In comparison with baseline methods presented in previous studies on SemEval Task 8, our approach—despite employing small-scale models (*sentence-transformers/all-MiniLM-L6-v2*) for latency optimization—achieves competitive accuracy. The synergy between hybrid search (fusing keyword and semantic signals) and a specialized re-ranking step allows the system to simultaneously broaden its search scope and maintain high precision in document ranking.

6 Conclusion

In this study, we developed and evaluated a RAG-based system for information retrieval in multi-turn conversations within the framework of SemEval-2026 Task 8.

Experimental results demonstrate that the hybrid search strategy (combining dense and sparse search) delivers superior performance compared to traditional methods, particularly when fine-tuned at the 30:30:30 configuration. Furthermore, the integration of a reranker model proved to be decisive in enhancing ranking precision, leading to a substantial increase in nDCG@10 and providing a more reliable context for Large Language Models. Despite persistent challenges in specialized domains such as Finance (FiQA), the proposed system exhibits robust generalization capabilities across various sectors, ranging from government to technology.

For future work, we aim to investigate domain-specific embedding models to improve performance in complex areas like finance. Additionally, we plan to research index optimization techniques to increase retrieval speed while maintaining high accuracy for real-time conversational systems.

Acknowledgement

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund. We thank the anonymous reviewers for their time and helpful suggestions that improved the quality of the paper.

References

Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. 2025. CORAL: Benchmarking multi-turn conversational retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1308–1330, Albuquerque, New Mexico. Association for Computational Linguistics.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*.

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. *arXiv 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. *arXiv preprint arXiv:2005.11401*.

Sara Rosenthal, Yannis Katsis, Vraj Shah, and Marina Danilevsky. 2025. Mtrageval: Evaluating multi-turn rag conversations. Available at: https://ibm.github.io/mt-rag-benchmark/MT_RAG_SemEval_Proposal.pdf.

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*, pages 2301–2305, Washington DC, USA. ACM.