TechSSN at SemEval-2026 Task 8: MTRAG Retrieval and Generation using Ensemble Re-encoders and Anchor Prompting

Anishka K, Anne Jacika J, Guruprakash K, Rajalakshmi Sivanaiah, S. Angel Deborah
Department of Computer Science and Engineering
SSN College of Engineering, Chennai, India
{anishka2310506, annejacika2310581, guruprakash2310495}@ssn.edu.in
{rajalakshmis, angeldeborahs}@ssn.edu.in

Abstract

This paper discusses the Retrieval-Augmented Generation (RAG) system submitted to the MTRAG-UN shared task on multi-turn conversational question answering. The paper describes the proposed solution for Task A (Document Retrieval) and Task C (Full RAG Pipeline), focusing on retrieval robustness and grounded response generation in complex English multi-turn dialogs. The proposed retrieval architecture uses a cascaded hybrid pipeline, which combines sparse retrieval (BM25) with dense bi-encoder models (BGE-base-en-v1.5 and E5-base), integrated via Reciprocal Rank Fusion and refined using a weighted ensemble of cross-encoders. For the generation part, the top-3 retrieved passages are injected into FLAN-T5-Large using an anchor-prompting strategy to output grounded faithful responses. Experimental results show that the proposed hybrid retrieval framework with multi-stage reranking significantly enhances passage selection, particularly for non-standalone conversational queries. Further analysis reveals persistent difficulties in handling underspecified and unanswerable questions, as well as an increased susceptibility to retrieval noise in later dialog turns.

1 Introduction

Retrieval-Augmented Generation (RAG) has become an inevitable component in the construction of Knowledge Based Conversational AI systems. By integrating external document retrieval with language generation, RAG frameworks aim to reduce hallucinations, enhance domain-specific knowledge utilization, and improve response traceability. Despite their effectiveness in single-turn tasks, RAG systems face increasing challenges in realistic conversational. Multi-turn interactions often involve user intent and contextual information distributed across successive turns, making query interpretation fundamentally context-dependent. Another challenge in the design of RAG systems is the ambiguity and unanswerability, implicitly associated with the queries from the user that surface from the available data.

The MTRAG-UN shared task seeks to systematically evaluate these challenges by providing a benchmark of 666 tasks across six domains (Rosenthal et al., 2026a; Katsis et al., 2025), explicitly targeting four critical phenomena: (1) Unanswerable questions, where no supporting evidence exists; (2) Underspecified questions, characterized by ambiguity and multiple plausible interpretations; (3) Non-standalone questions, which require prior conversational context for interpretation; and (4) Unclear responses, where users seek clarification or refinement of earlier system outputs.

Conversational RAG systems face two key challenges: implicit references that degrade retrieval and generation models prone to hallucinations. Instead of introducing new components, we evaluate established RAG techniques within multi-turn constraints. We focus on how standard RAG behaves in multi-turn contexts, using query rewriting for disambiguation and anchor prompting with structured abstention(Feng et al., 2024). This dual approach enables us to study the interaction between contextual retrieval and grounded response generation in multi-turn settings. The implementation of our system is available at: GitHub

2 Background

The MTRAG benchmark provides multi-turn conversational dialogs paired with domain-specific document corpora. Each task is a full conversation history that alternates between user and system with a domain identifier. Task A requires retrieving relevant pages, while Task C adds generating a faithful, grounded response to the user query.

Task A evaluates the retrieval quality using Recall@k and nDCG@k at $k \in 5, 10$ evaluated only for the 468 answerable questions and partially an-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2114–2119

swerable questions. Unanswerable are excluded from the scoring due to the unavailability of relevant passages to be retrieved. The performance of Task C is gauged through a set of three complementary metrics : RB_{llm} (LLM Judged Semantic similarity), RB_{agg} (Lexical ROUGE-L based overlap), and RL_F (RAGAS faithfulness in the retrieved passage). For unanswerable instances, IDK Judge measures the calibration of abstention of the system.

We participated in both Task A and Task C (Rosenthal et al., 2026b), covering four original MTRAG corpora. Two additional corpora introduced in MTRAG-UN, Banking and Telco, are outside the scope of our submission. Related works on Conversational Query Rewriting, dense retrieval (Zhao et al., 2024) and instruction tuned generation (Zhang et al., 2026) validates our design choices.

All four corpora present retrieval and generation challenges. FiQA is the most difficult; low lexical similarity between answers and source passages requires navigating indirect financial commentary. Conversely, IBM Cloud rewards the system with strong term-to-term matching. ClapNQ necessitates retrieving cohesive, multi-sentence passages for long factoid answers, while government documents require broad semantic coverage and complex canonical comprehension.

3 System Overview

The proposed system is a full sequential three stage pipeline: (1) Query rewriting, (2) Hybrid retrieval with Ensemble Reranking and (3) Grounded Response Generation. The first two stages constitute Task A, while the final stage corresponds to Task C. Figure 1 and 2 details the flow of the system.

3.1 Query Rewriting

The conversational history is transformed into a stand-alone query using Google Gemma-2-2B-IT(Team et al., 2024), a 2B parameter instruction-tuned model. The model was chosen based on the following three reasons:

- Strong adherence to structured prompts – Ensures accurate resolution of references while preserving named entities and original intent.
- Efficient deployment capability – Supports 4-bit NF4 quantization, enabling execution on a single T4 GPU with reduced memory footprint.
- Effective few-shot instruction following – Demonstrates strong performance without requiring additional fine-tuning.

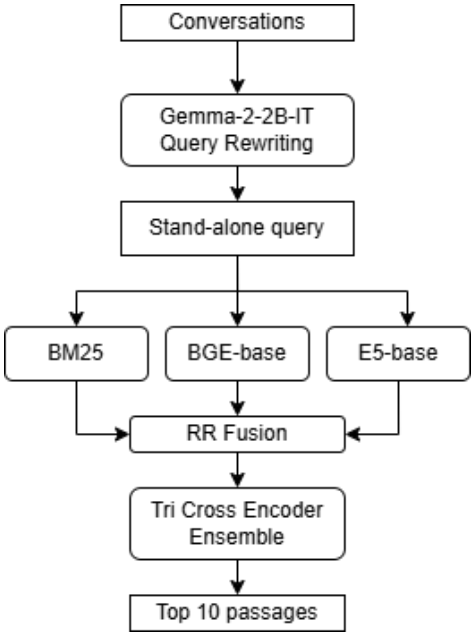

Figure 1: Architecture of the proposed framework for Task A.

3.2 Hybrid Retrieval and Ensemble Re-ranking

The rewritten query is processed by three parallel retrieval systems, whose ranked outputs are then consolidated using Reciprocal Rank Fusion (RRF) and subsequently refined through weighted cross-encoder re-ranking.

3.3 BM25 Lexical Retrieval

BM25 (rank-bm25) captures domain-specific terms and proper nouns that may be underrepresented in dense embeddings. The corpus is indexed using lowercase whitespace tokenization, making it effective for technical domains such as IBM Cloud and Government where lexical similarity provides strong adherance.(Ma et al., 2023)

3.4 Dense Retrieval (BGE-base-en-v1.5)

The system uses BGE-base-en-v1.5(Karpukhin et al., 2020), a 110M parameter bi-encoder for English retrieval, employing CLS pooling and cosine similarity with FAISS IndexFlatIP for efficient search. It performs strongly without requiring query prefixes.

3.5 Dense Retrieval (E5-base)

It is a 110M parameter bi-encoder trained with contrastive objective i.e., learn the relation through

Model	Weight	Reason
BAAI/bge-reranker-large	0.40	High individual nDCG on dev set
mixedbread-ai/mxbai-rerank-large-v1	0.35	Complements BGE on semantic queries
cross-encoder/ms-marco-MiniLM-L-12-v2	0.25	High throughput and fast for general purpose

Table 1: Tri-cross-encoder ensemble configuration and weights.

similarity and dissimilarity. It requires prefixes like "query: " and "passage: " due to its architectural requirement during training to directly promote the quality of retrieval. FAISS IndexFlatIP is used for indexing (helps in efficient retrieval) and similarity search(Zhan et al., 2021).

3.6 Reciprocal Rank Fusion

RRF aggregates rankings by summing reciprocal rank scores with a constant offset (k=60), ensuring robustness without explicit score normalization. The top K fused candidates are forwarded for reranking(Cormack et al., 2009).

3.7 Tri-Cross-Encoder Ensemble Re-ranking

The top 100 RRF candidates are reranked using three cross-encoders: BGE-large, MXBAI-base, and MS-MARCO-MiniLM. Table 1 describes the Min-Max normalized scores in the range [0,1], combined using weighted aggregation (0.40, 0.35, 0.25 respectively), where weights were assigned based on relative nDCG contributions observed during development experiments. The top K passages form the final output of Task A.

3.8 Grounded Response Generation Task C

Model: FLAN-T5-Large. The system uses Google/flan-t5-large, a 780M parameter model fine tuned on instructions following. FLAN-T5 was chosen for the following reasons: (1) Produces focused and bound outputs; (2) Operates comfortably with FP16 precision on a single TPU; (3) Fine tuning on instruction makes it responsive to strict source-grounding constraints in the prompt (Longpre et al., 2023). The model is loaded in FP-16 presicion with `device_map="auto" & low_cpu_mem_usage=True`

3.9 Selection with Cleaning

Although up to ten passages are retrieved, only the top three are provided to mitigate the "Lost in the Middle" effect(Liu et al., 2024). Retrieved passages are cleaned to remove URLs and headers, particularly important for cloud-domain corpora.

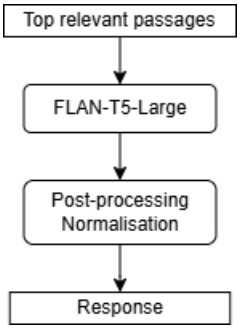

Figure 2: Architecture of the Generation framework for Task C.

3.10 Anchor Prompting

We employ anchor prompting to constrain generation. The model is instructed to answer strictly from the provided sources and return a fixed fallback statement ("I'm sorry, but I don't have the answer to your question.") if the answer is absent. The response begins with "Based on the sources," ensuring structural consistency and reducing hallucination(Noy and Musen, 2001).

3.11 Post-processing and Unanswerable Normalization

Generated outputs undergo prefix stripping to remove redundant labels. Low-confidence or ambiguous answers are normalized to a fixed fallback response (e.g., "I don't know the answer") to ensure consistency during evaluation by the IDK judge.

4 Experimental Setup

All experiments were run on Google Colab with T4 GPUs and Google Drive for persistent storage. GPU memory is actively managed between pipeline stages using `torch.cuda.empty_cache()` and `gc.collect()` to prevent out-of-memory errors.

4.1 Dataset & Preprocessing

The proposed system uses the official development split for tuning and evaluation, avoiding task-specific fine-tuning to maintain a zero-shot setting. This evaluates how pretrained models generalize to multi-turn conversations. Retrieval models (BM25,

Parameter	Value	Rationale
max_new_tokens	60	Avoid rambling
num_beams	3	Stay close to source text
no_repeat_ngram_size	3	Prevent repetition
length_penalty	1.0	Avoid length-based bias
early_stopping	True	Stop at first EOS, prevent padding

Table 2: FLAN-T5 generation hyperparameters.

BGE-base, E5-base), ensemble weights, and configurations were systematically tuned on the development set using nDCG@5. Multiple retrieval combinations, reranker weights, and top-K selections were evaluated; the final setup reflects the best-performing configuration on the development split. The FLAN-T5 module was tested on the development set for qualitative tuning and the official test for submission.

BGE and E5 corpus embeddings were precomputed offline per domain and serialized as pickle files. At execution, FAISS IndexFlatIP and BM25 indices are reconstructed from these stored files and passage texts.

4.2 Evaluation Metrics

Task A: Recall@5, Recall@10, nDCG@5, nDCG@10, evaluated on answerable and partially answerable instances only.
Task C: RB_{llm} (LLM Judged Semantic similarity), RB_{agg} (Lexical ROUGE-L Algorithmic) and RL_F (RAGAS faithfulness) and IDK (Abstention accuracy) and their harmonic mean, following recent RAG evaluation frameworks. (Fadnis et al., 2025)

4.3 Implementation Details

Table 2 shows the Hyperparameter specifics and the rationale behind the choice with respect to FLAN-T5 model.

- `max_new_tokens` : Controls the generation length
- `num_beams` : Number of candidates to choose in each round instead of greedily choosing the top one
- `no_repeat_ngram_size` : Prevent the model from repeating the same ngram
- `length_penalty`: A length controller in which values > 1 promote longer sequences, values < 1 encourage brevity, and a value of 1 maintains length neutrality.
- `early_stopping` : When all beams reach EOS token, the mode stops generation

Parameters were tuned on the development set to balance quality, faithfulness, and stability.

5 Results and Discussion

5.1 Task A: Retrieval

The system achieved a nDCG@5 of 0.5337, clearly outperforming the strongest baseline (0.4795) and ranked 7th out of 38 submissions. Results confirm that combining lexical and dense retrieval with RRF and cross-encoder ensemble re-ranking provides strong semantic coverage and robust ranking performance.

5.2 Task C: Generation

The model obtained a harmonic mean of 0.3198 (RLF = 0.6011) that ranked 26 out of 29, indicating strong factual grounding through anchor prompting. However, lower semantic similarity scores reflect limitations in capturing nuanced responses compared to larger models. Error analysis shows that retrieval noise and over-conservative abstention significantly affect performance, leading to incorrect grounding and excessive 'I don't know' responses. This reflects limitations of FLAN-T5 in multi-turn grounding and contextual consistency.

5.3 Performance by Answerability and IDK Analysis

Performance dropped sharply for underspecified queries due to weak abstention calibration, while nearly 30% of answerable questions were incorrectly normalized to IDK. In general, 166 out of 507 responses were marked as IDK, revealing a miscalibration between grounding and uncertainty control.

5.4 Performance by Domain

Table 3. describes the performance of the system in the domain, where IBM Cloud performs best due to the strong lexical similarity captured by FLAN-T5.FiQA has the highest RLF (0.6645) but weaker alignment with analytical financial answers, while ClapNQ scores lower due to the 60-token limit on its long responses.

Domain	N	RB_{agg}	RL_F	RB_{llm}	Avg. Length (chars)
IBM Cloud	131	0.2663	0.5588	0.2905	55.8
FiQA	77	0.2187	0.6645	0.2500	64.8
ClapNQ	142	0.1893	0.4437	0.2377	45.7
Govt	157	0.1734	0.4554	0.1803	86.6

Table 3: Performance of the system across domains.

Method	Recall@5	nDCG@5
BM25	0.2441	0.1274
Hybrid-1 (BM25 + BGE-base-en-v1.5)	0.4879	0.2958
Hybrid-1 + Query Rewriting	0.5937	0.3561
Hybrid-2 (BM25 + BGE-base-en-v1.5 + E5-base) + Query Rewriting	**0.7167**	**0.4573**

Table 4: Impact of retrieval components on development set performance.

Configuration	HM	RB_{alg}	RL_F	RB_{llm}
FLAN-T5-Base, greedy, no anchor, 512 tok (prototype)	∼0.21	∼0.16	∼0.42	∼0.18
FLAN-T5-Large, greedy, no anchor, 512 tok	∼0.26	∼0.19	∼0.50	∼0.22
FLAN-T5-Large, beam=3, no anchor, 1024 tok	∼0.28	∼0.21	∼0.55	∼0.25
FLAN-T5-Large, beam=3, anchor prompt, 1024 tok	∼0.30	∼0.23	∼0.58	∼0.27
FLAN-T5-Large + all improvements (submitted)	**0.3198**	**0.2444**	**0.6011**	**0.2759**

Table 5: Estimated performance comparison across different FLAN-T5 configurations and decoding strategies.

5.5 Abalation Study

Incremental ablation on the development set elaborated in Table 4 underscores the necessity of semantic components. The poor performance of the BM25 baseline confirms the limitations of lexical matching for conversational queries. Performance improves significantly with BGE-base-en-v1.5 (Hybrid-1), and further with query rewriting, which aids multi-turn disambiguation. An E5-base dual-encoder ensemble (Hybrid-2) performs best, showing complementary semantic coverage. Query rewriting boosts recall by resolving multi-turn references, while dense retrieval and reranking enhance semantic coverage and top-k precision to raise nDCG. Performance remains highest on ClapNQ and lowest on FiQA.

The proposed system's Task C baseline uses FLAN-T5-base (250M) with greedy decoding and no anchor prompting. This initial setup suffered from repetition (Longpre et al., 2023), hallucinations, and unreliable "I don't know" (IDK) judging due to a lack of normalization. Furthermore, uncleaned metadata in Cloud and ClapNQ distracted the generator. Table 5 provides an indicative ablation of these components; while not exhaustive, results highlight the necessity of prompt anchoring and metadata stripping.

6 Conclusion

The MTRAG system for the MTRAG-UN shared task addresses both document retrieval and full conversational RAG generation. The results for Task A show that hybrid retrieval integrating BM25 with BGE-base-en-v1.5 and E5-base, combined with query rewriting, significantly improves passage selection in multi-turn settings. In Task C, the system achieved a harmonic mean of 0.3198, with a relatively strong factual grounding score ($RL_F = 0.6011$), indicating effective hallucination-control through anchor prompting. However, lower semantic alignment scores reflect the limitations of FLAN-T5-Large compared to larger frontier models. Findings emphasize the central role of retrieval-robustness and structured grounding in building reliable conversational RAG systems.

7 Limitations and Future Work

The proposed system faces limitations such as excessive abstention ("I don't know") degrading performance, necessitating more granular prediction; failure to abstain on 72.4% of underspecified queries. Despite strong factual grounding ($RL_F = 0.60$), FLAN-T5-Large showed low semantic alignment ($RB_{llm} = 0.28$), struggling with referential nuances. Finally, computational constraints prevented evaluating larger architectures.

Future work will improve the abstention strategy, enhance response alignment, and explore more capable retrieval and generation models. Further refinements include query understanding mechanisms to better handle ambiguous and underspecified questions.

References

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Kshitij P Fadnis, Siva Sankalp Patel, Odellia Boni, Yannis Katsis, Sara Rosenthal, Benjamin Sznajder, and Marina Danilevsky. 2025. InspectorRAGet: An introspection platform for RAG evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 125–134, Albuquerque, New Mexico. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690.

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, Online. 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. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and 1 others. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International conference on machine learning*, pages 22631–22648. PMLR.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Natalya Fridman Noy and Mark A Musen. 2001. Anchor-prompt: Using non-local context for semantic matching. In *Ois@ ijcai*.

Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026a. Mtrag-un: A benchmark for open challenges in multi-turn rag conversations. *Preprint*, arXiv:2602.23184.

Sara Rosenthal, Vraj Shah, Yannis Katsis, and Marina Danilevsky. 2026b. Semeval-2026 task 8: Mtrageval: Evaluating multi-turn rag conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1503–1512.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and 1 others. 2026. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.