uir_cis at SemEval-2026 Task 8: A Unified Lightweight Pipeline for Multi-Turn RAG Evaluation

Jiaqi Zhang, Wenbin Duan, Yingqi Zhang, Yan Li, Binyang Li*
University of International Relations
{zjq,duan_wb,zyqzyqzyqzyq2021,liyan, byli}@uir.edu.cn

Abstract

This paper describes a system for SemEval-2026 Task 8 (MTRAGEval), covering Subtask A (retrieval) and Subtask B (generation) (Rosenthal et al., 2026b). We present a replicable multi-turn RAG pipeline design that runs entirely with open-weight models on a single GPU without relying on proprietary APIs, combining (1) query rewriting with Qwen2.5-7B-Instruct, (2) dense retrieval with BGE-M3, and (3) cross-encoder reranking with BGE-Reranker-v2-M3. For generation it reuses Qwen2.5-7B-Instruct with strict grounding instructions. On the official test set, the retrieval component achieves nDCG@5 of 0.4422 and the generation component achieves a harmonic mean score of 0.5664. We provide an ablation study quantifying the contributions of rewriting and reranking, and an error analysis guided by the organizers' analytics and answerability classes.

1 Introduction

Retrieval-augmented generation (RAG) underpins many practical assistants, but real deployments are typically *multi-turn* and include context-dependent user requests. Later turns may be underspecified and require integrating conversational context to retrieve correct evidence and produce grounded responses. SemEval-2026 Task 8 (MTRAGEval) evaluates multi-turn RAG across multiple domains with both retrieval (Subtask A) and grounded answer generation (Subtask B) (Rosenthal et al., 2026b; Katsis et al., 2025; Rosenthal et al., 2026a).

The study targets a *replicable, compute-efficient* system that runs entirely with open-weight models on a single GPU without relying on proprietary APIs under the official evaluation protocol. The approach uses a single open-weight 7B instruction-tuned model (Qwen2.5-7B-Instruct) for both query rewriting and answer generation, paired with strong off-the-shelf retrieval components: BGE-M3 embeddings (Chen et al., 2024) and a BGE cross-encoder reranker from the FlagEmbedding ecosystem (FlagOpen, 2024). The design avoids reliance on extremely large proprietary models and emphasizes end-to-end replicability and implementation transparency.

A unified formulation for Subtasks A and B is methodologically coherent because retrieval and generation operate over the same conversational state and evidence pool. Although the subtasks are evaluated separately, they share representations and preprocessing, enabling controlled reuse of dialogue formatting and rewriting components across both subtasks. The modular pipeline permits clean ablations of rewriting and reranking effects without changing downstream generation settings. This structure supports analysis of how retrieval specificity propagates to grounded answers.

Contributions include: (1) a detailed system description sufficient for reimplementation, (2) an ablation study isolating the effects of query rewriting and reranking under the organizers' evaluation protocol, and (3) an error analysis focusing on answerability and evidence grounding, informed by the organizers' analytics and InspectorRAGet-style inspection (Fadnis et al., 2025).

2 Background

2.1 Task and Data

SemEval-2026 Task 8 (MTRAGEval) evaluates multi-turn retrieval-augmented systems across multiple domains and conversational settings. Each instance provides a dialogue history and a target user query, and is associated with one of four domain collections: financial QA (FIQA), cloud support QA (CLOUD), open-domain QA (CLAPNQ), and government information QA (GOVT) (Katsis et al., 2025). All task data are in English.

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3143–3148

Subtask A (Retrieval). Given the conversation context, systems return a ranked list of passages from the corresponding domain corpus. A submission consists of the top-k passage identifiers with scores, and is evaluated with standard retrieval metrics; we focus on the official nDCG@5.

Subtask B (Generation). Given the conversation context *and* a set of reference documents, systems generate an answer grounded in the provided evidence and output "I don't know" when the documents do not contain sufficient information (Katsis et al., 2025; Rosenthal et al., 2026a). The test set includes an answerability phenomenon termed *underspecified*, which penalizes systems that answer confidently despite insufficient evidence (Rosenthal et al., 2026a).

2.2 Related Work

Conversational retrieval and query rewriting. Multi-turn conversational retrieval presents unique challenges because later turns are often context-dependent and may contain ambiguous references or underspecified queries. Unlike single-turn search, accurate retrieval frequently requires resolving conversational context into a standalone query. Prior work has shown that query reformulation or rewriting can improve retrieval for context-dependent turns by reducing ambiguity in conversational passage retrieval settings (Dalton et al., 2020; Voskarides et al., 2020; Lin et al., 2021). However, rewriting introduces an inherent trade-off: when queries contain domain-specific entities or numerical constraints, an LLM-based rewrite can drift from the original intent and harm retrieval accuracy. Our system treats rewriting as a modular component and explicitly analyzes its domain sensitivity, showing that rewriting can help when turns are underspecified but may hurt domains that require strict constraint preservation.

Dense retrieval and reranking. A common retrieval architecture is a two-stage pipeline: a fast retriever generates a candidate set, followed by a more accurate reranker that refines the top results. Bi-encoder dense retrieval provides efficient semantic matching but is limited by independent encoding, which can make it difficult to distinguish among highly similar candidates—precisely the regime that matters for small-cutoff metrics such as nDCG@5. Cross-encoder rerankers address this limitation by jointly modeling the query–passage interaction and often yield substantial gains in precision at top ranks (Karpukhin et al., 2020; Nguyen et al., 2016; Nogueira and Cho, 2019). Our approach follows this pattern using BGE-M3 embeddings (Chen et al., 2024) for dense retrieval and a BGE cross-encoder reranker (FlagOpen, 2024), and our ablations show reranking is a key driver of improvements across domains.

RAG and grounded evaluation. Retrieval-augmented generation (RAG) combines retrieval and generation, but ensuring that outputs are grounded in evidence remains challenging. Recent work emphasizes robust evaluation of grounding and answerability, including diagnosing hallucination and handling unanswerable queries (Lewis et al., 2020; Thorne et al., 2018; Es et al., 2024). Task 8 builds on MTRAG and provides analytics tooling (InspectorRAGet) to support fine-grained inspection of retrieval and generation behavior (Katsis et al., 2025; Fadnis et al., 2025). In this context, a central tension is balancing over-answering under weak or broad evidence against overly conservative abstention. Our system leverages strict grounding instructions and an explicit "I don't know" policy, and we use the task analytics to analyze recurring failure modes that connect back to upstream evidence selection and conversational query interpretation.

3 System Description

Figure 1 illustrates the unified pipeline. The system implements three major components: (1) query rewriting, (2) two-stage retrieval (dense retrieval + reranking), and (3) grounded generation. For Subtask A, the retrieval stage outputs the top-10 passages; for the official Subtask B submission, generation is conditioned on organizers-provided reference documents (no additional retrieval) with an explicit "I don't know" fallback.

3.1 Query Rewriting

Multi-turn user turns can be incomplete without the dialogue history. The current user turn is rewritten into a standalone query using Qwen2.5-7B-Instruct. The input prompt concatenates the dialogue history and the current user query, instructing the model to output *only* the rewritten query. Decoding uses greedy decoding (`do_sample=False`) with a maximum output length of 64 tokens. The rewritten query is used for retrieval in Subtask A.

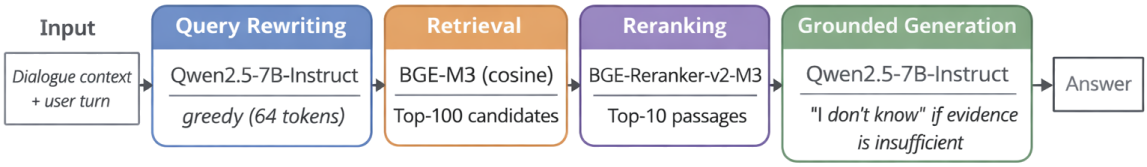

Figure 1: Unified pipeline for multi-turn retrieval and grounded generation: Subtask A outputs top-10 retrieved passages, while Subtask B generates from organizers-provided reference documents.

3.2 Two-Stage Retrieval: Dense Retrieval and Reranking

Dense retrieval. BGE-M3 is used as a bi-encoder to embed queries and passages. Following instruction-tuned embedding practice, we prefix each query with a short retrieval instruction ("Represent this sentence for searching relevant passages:"). Let $\mathbf{e}_q, \mathbf{e}_d \in \mathbb{R}^m$ denote the query and passage embeddings. Relevance is scored with cosine similarity:

$$s_{\text{dense}}(q, d) = \frac{\mathbf{e}_q^\top \mathbf{e}_d}{\|\mathbf{e}_q\| \, \|\mathbf{e}_d\|}. \quad (1)$$

The top-K candidates are retrieved by sorting $s_{\text{dense}}(q, d)$ (default K=100).

Cross-encoder reranking. The top-K candidates are reranked using BGE-reranker-v2-m3 (FlagEmbedding), which scores each query–passage pair with a cross-encoder:

$$s_{\text{rerank}}(q, d) = f_\theta([q; d]), \quad (2)$$

where $[q; d]$ concatenates the query text with the passage text (title + body). The top-k passages are selected after reranking (in this work, k=10). Dense retrieval returns K=100 candidates, reranking produces the top-10 list, and nDCG@5 is computed on the final ranking.

Ablation variants. For analysis, four retrieval variants are evaluated: (S1) dense-only, (S2) dense+rerank, (S3) rewrite+dense, and (S4) rewrite+dense+rerank (full system). All variants are evaluated with the organizers' official qrels and the same evaluation script.

3.3 Grounded Generation

For official Subtask B submission, Qwen2.5-7B-Instruct generates answers grounded in the organizers-provided reference documents (no additional retrieval). The prompt is formatted with a system instruction enforcing evidence-grounded answers and requiring "I don't know" when the evidence is insufficient, a list of provided reference documents, and the dialogue history. To control context length and memory usage, each reference document is truncated to 1200 characters. We use standard generation settings; full decoding hyperparameters are provided in Appendix A.2.

Algorithm 1 Pipeline for Subtask A retrieval and Subtask B grounded generation from provided documents.

Input:
Dialogue history H, current user turn u, domain corpus $\mathcal{D}$ (Subtask A)
Organizers-provided reference documents $\mathcal{R}$ (Subtask B)
$q \leftarrow \text{Rewrite}(H, u)$ (Qwen2.5-7B, greedy)
$C_K \leftarrow \text{DenseRetrieve}(q, \mathcal{D}, K{=}100)$ (BGE-M3)
$C_{10} \leftarrow \text{Rerank}(q, C_K, 10)$ (BGE reranker)
Output top-10 passages C_{10} (doc_id, score) for Subtask A
$a \leftarrow \text{Generate}(H, u, \mathcal{R})$ (Qwen2.5-7B, grounded)
(No additional retrieval in Subtask B)
return C_{10}, a

3.4 Pipeline Pseudocode

Algorithm 1 formalizes our end-to-end pipeline and highlights the shared structure between Subtask A and Subtask B. We first rewrite the multi-turn query into a standalone form, then perform dense retrieval followed by cross-encoder reranking to obtain the Subtask A retrieval output (top-10 passages). For official Subtask B submission, grounded generation consumes the organizers-provided reference documents with no additional retrieval. While our approach relies on standard components, its value lies in a unified, replicable design that can run with open-weight models on a single GPU, with fixed hyperparameters documented in the appendix, and supports transparent analysis of rewriting and reranking effects.

4 Experimental Setup

Hardware. All experiments run on a single NVIDIA A800 with a 20GB allocated memory slice.

Subtask	Metric	Score	Rank
A (Retrieval)	nDCG@5	0.4422	19/38
B (Generation)	Harmonic mean	0.5664	21/26

Table 1: Test results for the system as reported by the organizers (Rosenthal et al., 2026b).

Evaluation protocol. For Subtask A, we evaluate retrieval using the organizers' official qrels and their `pytrec_eval`-based script, reporting nDCG@5 as the primary metric. For Subtask B, generation is conditioned on the organizers-provided reference documents, and we report the official harmonic mean score (aggregating RB_agg, RL_F, and RB_llm) provided by the organizers.

Implementation notes. Dense retrieval uses cosine similarity over BGE-M3 embeddings; reranking uses FP16 inference with a batch size of 16. All ablation settings use the same corpora and a fixed local evaluation protocol to ensure controlled relative comparisons.

5 Results and Analysis

5.1 Official Results

Table 1 reports test results for the system on Subtasks A and B. On the official leaderboard, the system ranks 19/38 on Subtask A with nDCG@5 of 0.4422 and 21/26 on Subtask B with a harmonic mean of 0.5664 (Rosenthal et al., 2026b).

5.2 Ablation Study (Subtask A)

Table 2 reports domain-wise nDCG@5 and an overall weighted average (weighted by the number of queries per domain). Organizer-reported scores in Table 1 are the official ranking numbers. Table 2 reports local evaluation on the officially released test set; Overall is a query-count-weighted average across the four domains. Small differences versus the organizer-reported official test score in Table 1 can arise from weighted aggregation versus the organizers' aggregation. Reranking is the main driver of gains, improving all four domains relative to dense-only retrieval. Rewriting has mixed effects: it helps CLAPNQ but hurts FIQA, while CLOUD and GOVT are closer to neutral. Rewrite-only (S3) reduces FIQA performance (0.1399 vs. 0.2076 in S1). We hypothesize that FIQA queries are particularly sensitive to precise financial entities and numerical constraints (e.g., tickers, amounts, dates), and LLM-based rewriting can introduce paraphrastic noise or slight constraint drift that harms retrieval. In contrast, CLAPNQ benefits from rewriting, suggesting that rewriting is most helpful when conversational turns are underspecified and require context resolution. Overall, the full system performs best, indicating that reranking complements rewriting by improving ordering at small cutoffs and mitigating rewrite drift.

Setting	FIQA	CLOUD	CLAPNQ	GOVT	Overall
S1 (Dense Only)	0.2076	0.2742	0.3339	0.2409	0.2705
S2 (+ Rerank)	**0.3419**	0.4400	0.4292	0.4364	0.4210
S3 (+ Rewrite)	0.1399	0.3222	0.4730	0.3242	0.3374
S4 (Full System)	0.2885	**0.4410**	**0.4850**	**0.4943**	**0.4466**

Table 2: Ablation study of retrieval components across domains (nDCG@5). Overall is the weighted average across domains using the number of queries in each domain as weights.

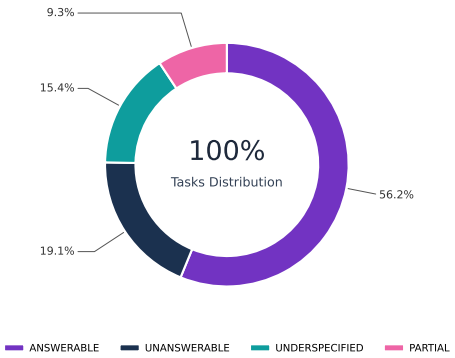

Figure 2: Distribution of answerability classes in the evaluation set.

5.3 Error Analysis (Subtask B)

Error analysis is guided by the official analytics categories. The focus is on the interaction between retrieval specificity and generation grounding. Multi-turn context increases sensitivity to rewrite errors that shift retrieval targets. Figure 2 shows the distribution of answerability classes in the evaluation set, highlighting the presence of underspecified and unanswerable queries that require careful IDK handling.

Two recurring failure modes are analyzed: over-answering under multi-document evidence and IDK responses driven by evidence mismatch.

Case 1: Over-answering / information over-aggregation under multi-document evidence. For a query on the evolution of the Prime Minister's role, provided reference documents span multiple national contexts, expanding the scope beyond the intended focus. The generator aggregates across these contexts, introducing an aggregation bias that

favors breadth over specificity. Evidence such as "appointed and dismissed Cabinet members" is amplified into a multi-point narrative. Evaluation penalizes verbosity, so the expanded scope reduces score despite grounded content. The failure reflects scope drift rather than hallucination.

Case 2: IDK answer despite provided documents (evidence mismatch). For a flu-vaccine eligibility query, the model abstains with "I don't know" under weak evidence. Provided passages are generic (e.g., vaccination is a "cheap and effective way" to prevent infections) and omit population categories. Given the mismatch, abstention is the correct behavior; the failure is upstream, caused by insufficient retrieval granularity. The case is not a hallucination but a retrieval specificity error that prevents grounded answering. Prioritizing passages with explicit eligibility groups would mitigate this failure mode.

Ethical Considerations

RAG systems can produce incorrect or over-aggregated answers, which may mislead users in domains such as finance or government. Our system mitigates this risk with strict grounding instructions and an explicit "I don't know" fallback when evidence is insufficient. The shared-task data and organizer attachments have redistribution restrictions; we do not redistribute the data or predictions.

6 Conclusion

A system is presented for SemEval-2026 Task 8, combining Qwen2.5-7B-based query rewriting and grounded generation with BGE-M3 dense retrieval and BGE cross-encoder reranking. Ablations show that reranking yields consistent gains across domains and that query rewriting provides complementary improvements when combined with reranking, while rewriting alone can be domain-sensitive. Both error cases point to a single bottleneck: generation quality is tightly bounded by evidence selection and control. When retrieval surfaces multiple loosely related passages, the generator tends to over-aggregate and produce verbose answers; when retrieval fails to supply specific supporting evidence, the system correctly abstains with an IDK response. Future work should therefore focus on tighter evidence filtering/selection and explicit answer length control to better align multi-turn RAG outputs with the evaluation protocol. Future work includes explicit answerability classification, improved history-aware rewriting, and stronger evidence attribution mechanisms to reduce over-answering.

Acknowledgments

We thank the SemEval-2026 Task 8 organizers for providing the MTRAGEval benchmark and evaluation infrastructure. This work was supported by the Beijing Natural Science Foundation (Grant number: 4262075) and the Research Funds for NSD Construction, University of International Relations (Grant numbers: 3262026T23).

References

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pages 2318–2335.

Jeff Dalton et al. 2020. TREC Conversational Assistance Track (CAST). In *Proceedings of the 29th Text REtrieval Conference (TREC)*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Kshitij Fadnis et al. 2025. An introspection platform for RAG evaluation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Demo Track)*.

FlagOpen. 2024. Flagembedding: Retrieval and reranking models (bge family). `https://github.com/FlagOpen/FlagEmbedding`.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, et al. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Jimmy Lin et al. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*.

Tri Nguyen, Mir Rosenberg, Xia Song, et al. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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. *arXiv 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.

James Thorne et al. 2018. The FEVER shared task: System evaluation of fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and Verification (FEVER)*.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

A Implementation Details

A.1 Query Rewriting Prompt

Exact rewrite prompt template.

```
System: Rewrite the current user
    question into a standalone search
    query.
Return only the rewritten query and
    nothing else.

Conversation History:
{HISTORY_TEXT}

Current Question:
{CURRENT_USER_TURN}
```

Dialogue history is concatenated in chronological order with explicit "User:"/"Assistant:" turn markers. If the concatenated history exceeds the model context budget, we keep the most recent turns (tail truncation) and always retain the current user turn.

A.2 Generation Prompt and Configuration

Our generation module uses Qwen2.5-7B-Instruct with a strict grounding prompt. For official Subtask B, input documents are the organizers-provided reference documents. Each reference document is truncated to 1200 characters before prompt assembly. This is implemented as a simple hard truncation heuristic to control context length and memory usage under the 20GB memory budget. Dialogue history is concatenated in chronological order with turn markers, and if history exceeds the context budget, the most recent turns are kept (tail truncation).

Exact generation prompt template.

```
System: You must answer strictly based
    on the provided documents.
If the answer cannot be found in the
    documents, respond with "I don't
    know."

Documents:
[Doc 1]
[Doc 2]
...

Conversation History:
User: ...
Assistant: ...

Current Question:
...
```

For decoding, we use max_new_tokens = 512, temperature = 0.7, and top_p = 0.9. All generation experiments use identical decoding settings to ensure comparability across runs.