DUTIR at SemEval-2026 Task 8: A Hybrid Retrieval and Faithfulness-Guarded Framework for Multi-Turn RAG

Ruiyang Jin, Yichong Chen, Liang Yang*
DUTIR Lab, Dalian University of Technology
{ruiyangjin, 15631076392}@mail.dlut.edu.cn
liang@dlut.edu.cn

Abstract

This paper describes the system submitted by **DUTIR_taskC** for SemEval-2026 Task 8: MTRAGEval (Task C). Multi-turn Retrieval-Augmented Generation (RAG) poses significant challenges in context tracking, retrieval precision, and hallucination mitigation. Our proposed system addresses these by employing a multi-stage pipeline consisting of: (1) LLM-based query rewriting (powered by **GPT-5.2**) to resolve conversational dependencies; (2) a hybrid retrieval module combining dense embeddings (BGE-M3) and sparse retrieval (BM25) with Reciprocal Rank Fusion (RRF); (3) a confidence-based answerability gating mechanism; and (4) a post-generation faithfulness guard. Experimental results on the blind test set show that our approach achieves a Composite Score of **0.5576**, ranking **4th out of 29** participating teams. Detailed analysis reveals that our system significantly outperforms strong baselines in faithfulness and successfully handles underspecified queries.

1 Introduction

Retrieval-Augmented Generation (RAG) has become the de-facto standard for knowledge-intensive NLP tasks (Lewis et al., 2020). This paper describes our system for SemEval-2026 Task 8: MTRAGEval (Rosenthal et al., 2026b), which focuses on evaluating RAG systems in English multi-turn conversational settings using the MTRAG-UN benchmark (Rosenthal et al., 2026a). Extending RAG to multi-turn conversations introduces critical complexities: user queries are often incomplete, exhibiting frequent anaphora (pronoun references) and ellipsis (omitted subjects) dependent on dialogue history. Furthermore, as conversations evolve, topic drifts compound the difficulty of fetching relevant historical context, making faithfulness maintenance over long, multi-document reasoning steps exceptionally difficult.

To address these challenges comprehensively, we present a robust, multi-stage framework for Task C. Our core contributions and main strategies are:

1. **Context-Aware Query Reformulation**: We leverage the advanced reasoning capabilities of GPT-5.2 to transform multi-turn dialogue history into standalone, retrieval-friendly queries, effectively bridging the semantic gap between conversational utterances and formal documents.

2. **Hybrid Retrieval & Reranking**: We leverage both semantic and keyword matching, refined by a cross-encoder reranker, to maximize recall across diverse domains.

3. **Faithfulness Guard**: We introduce a novel post-processing module that audits generated answers against retrieved contexts to mitigate hallucinations and calibrate the tone for unanswerable questions.

Experimental results on the blind test set show that our approach achieves a Composite Score of **0.5576**, ranking **4th out of 29** participating teams. While our rigorous gating minimizes hallucinations, qualitative analysis reveals occasional over-conservatism, leading to unwarranted "I Don't Know" (IDK) rejections for partially answerable queries.

2 Background

The MTRAGEval Task C requires participants to build a comprehensive multi-turn RAG pipeline. Unlike single-turn QA, this task evaluates the system as a conversational agent that continuously grounds responses in external knowledge across turns, demanding temporal context awareness.

* Corresponding author.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 328–332

Task Setup and Input/Output: At each turn, the system receives a *dialogue history* (previous user queries and system responses) and a *current user query*. The expected output consists of two parts: (1) a set of retrieved *contexts* (up to 5 passages) and (2) a generated *answer* strictly grounded in the retrieved contexts. For example, if the dialogue history discusses deep learning frameworks and the user's current query is *"How do I install it?"*, the system must first resolve the coreference ("it" refers to PyTorch), retrieve relevant official installation documentation, and generate a step-by-step answer based solely on those passages.

Datasets and Domain Shift: The task utilizes passage-level corpora comprising English documents spanning diverse domains, including finance (FiQA), government policies (Govt), cloud computing infrastructure, and general Wikipedia text (ClapNQ). Given this multi-domain nature, user queries often suffer from severe vocabulary mismatch compared to the formal corpora. Off-the-shelf retrievers frequently fail to align casual conversational language with dense domain-specific jargon. This necessitates intermediate query resolution and rigorous post-generation verification, inspiring the design of our query rewriter and faithfulness guard.

3 System Description

Our system architecture is illustrated in Figure 1. The pipeline consists of offline indexing and an online inference workflow.

3.1 Offline Indexing

We utilize the provided passage-level corpora. To enhance retrieval performance, we employ a **Hybrid Indexing Strategy**:

- **Dense Index**: We use BAAI/bge-m3 (Chen et al., 2024) to generate embeddings. Following the model's best practices, we prepend the instruction "passage: " to texts to align with the model's instruction-tuning objectives. Crucially, we concatenate the document title with the passage content during indexing to enrich semantic representation and preserve global document context.

- **Sparse Index**: We build a BM25 index (Robertson and Zaragoza, 2009) to capture exact keyword matches, which excels at capturing rare entities, acronyms, or technical terms often missed by dense models.

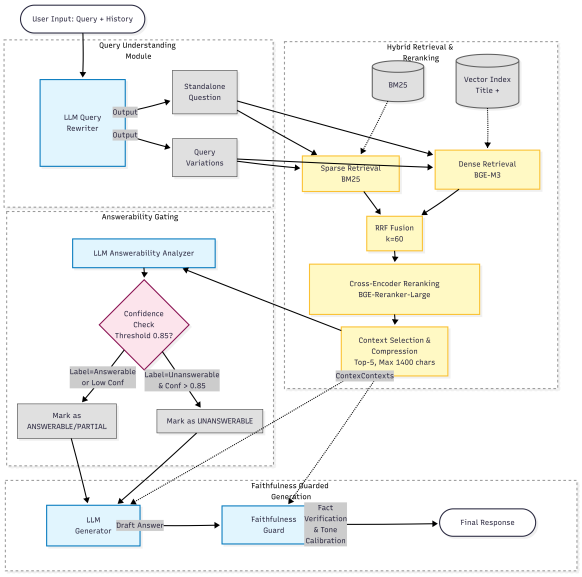

Figure 1: **System Architecture**. The pipeline features a GPT-5.2-powered query rewriter, hybrid retrieval (Dense+Sparse), and a confidence-aware faithfulness guard.

3.2 Multi-Turn Query Understanding

Directly using the raw user query often leads to retrieval failure due to missing subjects, predicates, or coreferences. We employ an LLM-based **Query Rewriter** that takes the conversation history and the raw query as input. Beyond simple coreference resolution, the rewriter addresses implicit topic continuation. The rewriter reconstructs omitted context and generates a standalone question to maximize downstream retrieval coverage.

3.3 Hybrid Retrieval and Reranking

For each rewritten query, we perform parallel retrieval to exploit the complementary strengths of our indexes:

- **Retrieval**: We retrieve top-k_{dense} documents using the vector index and top-k_{sparse} documents using BM25.

- **Fusion**: Results are merged using **Reciprocal Rank Fusion (RRF)** (Cormack et al., 2009) with $k = 60$. RRF effectively normalizes the disparate scoring scales between dense similarities and sparse TF-IDF scores.

- **Reranking**: The top-120 fused results are rescored using a Cross-Encoder (BAAI/bge-reranker-large), which computes full-attention

interactions between the query and the documents, capturing fine-grained relevance that bi-encoders miss.

- **Context Selection**: We select the top-5 documents and compress the total length to 1,400 characters to fit the generator's optimal context window.

3.4 Answerability Gating and Generation

A fundamental challenge in MTRAGEval is mitigating hallucinations when queries are unanswerable based on the retrieved texts. We implement a **Confidence-based Gating Mechanism**. The system classifies the query into *ANSWERABLE*, *PARTIAL*, or *UNANSWERABLE* via a zero-shot prompt. Unlike standard discrete classification, we instruct the LLM to output a confidence probability and enforce a strict threshold ($\tau = 0.85$). If predicted *UNANSWERABLE* with confidence $< \tau$, it downgrades to *PARTIAL* to encourage a safe answer attempt.

3.5 Faithfulness Guard

To further enforce strict adherence to the context, a secondary LLM module performs **Fact Verification** (removing any unsupported claims generated by the draft) and **Tone Calibration**. Tone calibration ensures that for *PARTIAL* labels, the system explicitly hedges its response (e.g., stating what information is missing).

3.6 System Workflow Walkthrough

To concretely illustrate our system's pipeline in action, consider a scenario where the dialogue history discusses the IBM Watson Assistant web chat, and the user inputs a highly fragmented query: *"Add user identity information"*. First, our **Query Rewriter** resolves the implicit context, generating a standalone query: *"How to add user identity information to the IBM Watson Assistant web chat integration?"* During Hybrid Retrieval, RRF and reranking surface the exact API documentation detailing the `updateUserID()` method. The **Gating Mechanism** then confidently classifies the query as *ANSWERABLE*, prompting the generator to draft a response based strictly on the retrieved documents. Finally, the **Faithfulness Guard** verifies the draft, ensuring no invalid parameters are hallucinated. In the final evaluation, this pipeline execution yielded a highly faithful output with an *RL_F* score of 0.90.

4 Experimental Setup

4.1 Data and Models

We used the provided SemEval-2026 Task 8 datasets and evaluation framework (Katsis et al., 2025), derived from the MTRAG-UN benchmark (Rosenthal et al., 2026a), consisting of **English** multi-turn dialogue histories and passage corpora.

- **Embedding & Reranker**: BAAI/bge-m3 and BAAI/bge-reranker-large.
- **LLM**: We employed **GPT-5.2** (OpenAI) for all generative components (rewriting, generation, guarding) due to its superior zero-shot reasoning and robust instruction-following.

4.2 Implementation Details

Our framework is implemented using PyTorch (Paszke et al., 2019) and FAISS (Johnson et al., 2019). Key retrieval hyperparameters are set as follows: BM25 and Dense Top-k at 200, Post-Rerank Top-k at 80, and a final context size of $N = 5$. To govern the LLM's behavior during the gating and guarding stages, we employ 3-shot system prompts mapping ambiguous dialogue scenarios to appropriate IDK (I Don't Know) or partial labels.

A critical parameter in this gating mechanism is the IDK confidence threshold (τ). Because this confidence score is generated directly by the LLM via prompting rather than derived from normalized token logits, we empirically tuned it on the official Task C development set. Evaluating τ across a range of 0.70 to 0.95 revealed a stark trade-off: lower thresholds ($\tau < 0.80$) failed to adequately suppress parametric hallucinations (degrading *RL_F* scores), whereas excessively high thresholds ($\tau > 0.90$) induced systemic over-conservatism, causing the model to erroneously reject partially answerable queries and severely penalizing recall metrics. Consequently, $\tau = 0.85$ was established as it achieves the optimal equilibrium between strict contextual faithfulness and conversational helpfulness.

5 Results and Analysis

5.1 Leaderboard Performance

Table 1 presents the performance of our system on the Task C blind test set. The primary evaluation metric is a Composite Score calculated as the harmonic mean of *RB_agg*, *RL_F*, and *RB_llm*.

Our system achieved a **Composite Score of 0.5576**, securing the **4th rank among 29 teams**.

Notably, our approach surpassed the provided strong baseline (*qwen-30b-a3b-thinking*) by a margin of 2.1 points, validating the effectiveness of our specialized RAG pipeline in highly constrained conversational settings.

Table 1: Leaderboard results for Task C (RAG).

System	Composite	RB_agg	RL_F	RB_llm
Top-1 System	0.5861	-	-	-
DUTIR	**0.5576**	0.4073	**0.7123**	0.6575
Baseline (Qwen)	0.5366	-	-	-

5.2 Metric Divergence Analysis

As shown in Table 1, we observe a significant divergence between our *RB_agg* (0.4073) and *RL_F* (0.7123). Based on the provided evaluation scripts (Katsis et al., 2025), *RB_agg* is calculated as the harmonic mean of ROUGE-L, BERTScore Recall, and Extractiveness. Generative models like GPT-5.2 tend to produce highly abstractive and fluent responses, which naturally diverge from extractive gold-standard references. This inherent characteristic explains the depressed *RB_agg* scores, which heavily penalize correct answers that fail to exact-match the reference string.

In contrast, the exceptionally high *RL_F* score (0.7123)—which relies on the RAGAS Faithfulness metric to evaluate whether the generated response is strictly grounded in the retrieved contexts—proves that our **Faithfulness Guard** effectively suppressed hallucinations. Furthermore, the robust *RB_llm* score (0.6575), an LLM-as-a-judge metric evaluated against the gold targets, confirms that our overall retrieval and generation pipeline produced highly accurate and helpful answers. This profile is visualized in Figure 2.

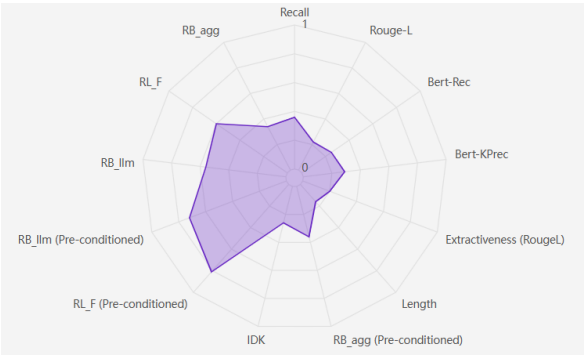

Figure 2: Performance breakdown across different metrics, visualized using InspectorRAGet.

5.3 Handling Underspecified Queries

According to the post-evaluation statistical analysis, the test set contained a significant portion of "Underspecified" queries (approx. 28.4%). These queries lack sufficient context for a definitive answer even after retrieval. The evaluation framework implements rigorous IDK conditioning: systems correctly identifying unanswerable contexts receive a perfect metric score (1.0), whereas incorrect hallucinated attempts yield zero. Our system exhibited an IDK rate of 0.26 on the test set. This close alignment proves that our **Answerability Gating Mechanism** (with $\tau = 0.85$) effectively identified ambiguous contexts, correctly withholding answers to avoid severe penalties.

5.4 Error Analysis

To better understand our system's limitations and identify areas for future improvement, we performed a deep qualitative analysis on the failure cases identified in the final evaluation report. We categorize the primary modes of error below, with detailed examples provided in Appendix A (Table 2). These errors prominently highlight the inherent tension between strict factual adherence and conversational helpfulness.

Over-Conservatism (The Strict Gating Dilemma): Our confidence threshold ($\tau = 0.85$) acts as a double-edged sword. While it effectively reduces hallucinations, it occasionally leads to unwarranted full IDK rejections. As shown in the first example of Table 2, when asked about US option trades, the system correctly deduced the premise of the question based on the retrieved context. However, because the exact, specific names of the exchanges were truncated or missing in the top retrieved chunks, the system conservatively refused to answer entirely, yielding an IDK score of 1.00. This strictness caused the system to miss partial credit that a lenient agent would have earned via a partial explanation.

Parametric Knowledge Hallucination (Memory Interference): While the Faithfulness Guard corrects most unsupported statements, some subtle hallucinations still bypass it, particularly during highly open-ended or enumerative queries. In the second example (Table 2), the user asked for *"any other terminology"* they should know. The LLM, triggered by the financial context of the prompt, relied on its pre-trained parametric memory to generate a plausible list of financial terms (e.g.,

NPV, DCF, PEG Ratio) that were objectively correct in the real world but completely absent from the retrieved texts. While this response is contextually highly helpful to a real user, it severely violates strict RAG faithfulness constraints. This phenomenon demonstrates how powerful LLMs can inadvertently prioritize helpfulness over strict grounding, resulting in severely penalized *RL_F* scores.

6 Conclusion

In this paper, we presented a hybrid RAG framework for SemEval-2026 Task 8. By combining context-aware query rewriting, hybrid multi-index retrieval, and a confidence-based faithfulness guard, our system achieved balanced and highly faithful performance, ranking 4th overall. Our deep error analysis highlights the critical importance and inherent difficulty of precise answerability detection in multi-turn scenarios, especially when balancing parametric helpfulness with extractive strictness. Future work explores dynamic thresholding based on query complexity, alongside fine-tuning lightweight, local evaluators to replace broad-purpose LLMs for controllable faithfulness guarding.

Acknowledgments

This work is supported by the Information Retrieval Laboratory at Dalian University of Technology (DUTIR Lab).

References

Jianlv Chen, Shitao Wang, Jianlin Liu, Chen Gao, Zheng Liu, and 1 others. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. RAG: Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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.

A Error Analysis Cases

Table 2 provides concrete examples of the failure modes discussed in Section 5.4. These cases highlight the trade-offs between rigorous faithfulness gating and helpfulness in complex multi-turn contexts.

Table 2: System failure modes identified during post-evaluation.

Type 1: Over-Conservatism (Strict Gating)
Query: *"Which one in the US trades options?"* **Context:** Mentions options trade on dedicated exchanges, but specific names were truncated. **Output:** *"The specific names... are not provided..."* **Impact:** Triggered full IDK rejection instead of a partial answer. **IDK: 1.00**
Type 2: Parametric Hallucination (Low Faithfulness)
Query: *"Any other terminology you think I should know?"* **Context:** Definitions for NAV, P/E ratio, etc. **Output:** Listed ungrounded terms like *NPV, DCF*, and *PEG Ratio* based purely on LLM's internal memory. **Impact:** Severely penalized faithfulness score. **RL_F: 0.25**