JCT at SemEval-2026 Task 8: Resource-Efficient Multi-Turn RAG via Nano-LLM Rewriting and Hybrid Reranking

Tal Farhan and Chaya Liebeskind
Jerusalem College of Technology
21 Havaad Haleumi St., 91160
Jerusalem, Israel
{talsh838, liebchaya}@gmail.com

Abstract

This paper describes our system submission for SemEval-2026 Task 8 (MTRAGEval), focusing on multi-turn Retrieval-Augmented Generation (RAG). Conversational queries often suffer from contextual ambiguity, rendering standard retrieval methods ineffective. We propose a highly resource-efficient pipeline that decouples query understanding from retrieval using a 1.5B parameter Nano-LLM (Qwen) for query rewriting, followed by parallel hybrid retrieval (Qdrant) and Cross-Encoder reranking. During internal development, our optimized system achieved an nDCG@5 score of 0.1991 on answerable queries, outperforming the official BM25 baseline (nDCG@5 of 0.18). On the official blind test set, the system achieved a score of 0.1744. While our absolute performance trails behind baselines utilizing massive 20B parameter models, our work establishes a crucial baseline for extreme resource efficiency in conversational RAG. We provide a comprehensive error analysis detailing the impact of domain shifts, retrieval funnels, and we conduct a qualitative analysis on the organizers' surprise "Underspecified" class to highlight the vulnerabilities of generative query rewriting.

1 Introduction

Retrieval-Augmented Generation (RAG) is essential for grounding Large Language Models (LLMs) in external knowledge, mitigating hallucinations. However, in multi-turn conversational settings, user queries are frequently ambiguous, anaphoric, or dependent on previous turns (e.g., "How does it compare to the previous version?").

In this work, we present our submission to SemEval-2026 Task 8 (Katsis et al., 2025; Rosenthal et al., 2026b). We implemented a multi-stage RAG pipeline that explicitly resolves coreferences and handles topic shifts before initiating retrieval. Our strategy emphasizes computational efficiency: utilizing a Nano-LLM (1.5B parameters) for query rewriting and open-weight embedding models. We demonstrate that while massive, proprietary models yield superior absolute accuracy, carefully engineered local pipelines can effectively navigate complex conversational dependencies with a fraction of the computational cost. The full implementation of our system is publicly available at https://github.com/TalFarhan/mt-rag-task-a.

2 Background and Related Work

The task of Conversational Information Seeking (CIS) extends traditional ad-hoc retrieval by introducing context dependency across conversational turns.

Query Rewriting in CIS: To bridge the gap between conversational context and dense retrievers, Query Rewriting (QR) has become a standard approach (Dalton et al., 2020). Recent approaches leverage LLMs to rewrite queries into standalone formats by resolving coreferences and omitting conversational filler (Yu et al., 2020). However, relying on massive, proprietary models introduces unacceptable latency for real-time applications. In our work, we explore the efficacy of a Nano-LLM (1.5B parameters) to perform this task locally, proving that smaller models can achieve high precision when properly prompted.

Hybrid Retrieval & Reranking: Dense retrievers often struggle with out-of-vocabulary terms, serial numbers, and domain-specific jargon. Hybrid retrieval strategies, which fuse sparse lexical signals (e.g., BM25) with dense embeddings using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), have shown superior robustness. This initial retrieval is typically followed by a Cross-Encoder reranking stage (Nogueira and Cho, 2019), which models the deep, token-level interaction between the query and document, acting as a highly precise but computationally expensive filter.

The Answerability Challenge in RAG: A crit-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1326–1331

ical limitation of traditional RAG pipelines is the implicit assumption that the target corpus actually contains the answer to the user's query (Rajpurkar et al., 2018). While much research has focused on improving retrieval accuracy for answerable questions, unanswerable or underspecified queries pose a unique and severe threat to dense retrieval systems. Without an explicit abstention mechanism, the system will inevitably retrieve the "nearest" semantic document, creating hard false-positives that mislead downstream generation models.

3 System Architecture

Our system architecture comprises three main components designed to handle conversational ambiguity and retrieve highly relevant context.

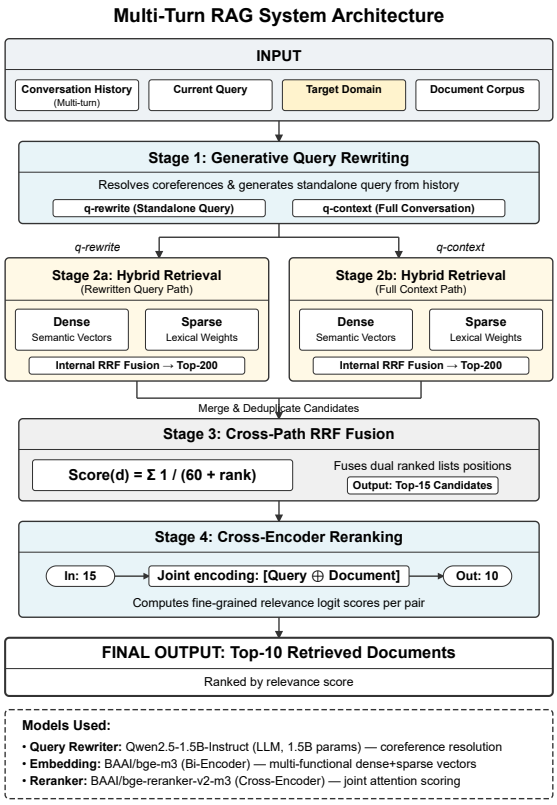

Figure 1: Overview of the proposed Multi-Turn RAG pipeline. The system performs generative query rewriting to resolve context dependencies, followed by parallel hybrid retrieval over two query representations, cross-path RRF fusion, and final cross-encoder reranking.

3.1 State Management and Dual-Query Generation

We implemented an Intent Router to maintain a global state dictionary mapping sequence IDs to conversational histories. To address multi-turn dependencies and contextual ambiguity, this module processes the current conversation turn to construct two distinct query representations:

- **LLM-Rewritten Query ($q_{rewrite}$):** We integrated the Qwen 2.5 1.5B Instruct LLM (quantized to 4-bit) as a query rewriter. It parses the chat history and the latest utterance to generate a self-contained, standalone query.
- **Full Contextual Query ($q_{context}$):** In parallel, the system constructs a raw query by concatenating the entire conversational history along with the latest user utterance into a single, continuous text string.

Both $q_{rewrite}$ and $q_{context}$ are subsequently utilized as independent inputs for the downstream parallel retrieval stage.

3.2 Parallel Hybrid Retrieval and Fusion

We indexed the provided corpus (Rosenthal et al., 2026a) using the `BAAI/bge-m3` model, generating both dense and sparse embeddings stored in Qdrant. During inference, we execute two parallel hybrid searches: one using the rewritten query, and one using the full raw conversational context. We fuse these streams using Reciprocal Rank Fusion (RRF):

$$Score(d) = \sum_{r \in R} \frac{1}{k + rank(r, d)} \quad (1)$$

Where d represents the candidate document being evaluated, and R is the set of all parallel retrieval streams (i.e., the dense and sparse ranking lists generated from both $q_{rewrite}$ and $q_{context}$). For each retrieval stream $r \in R$, $rank(r, d)$ denotes the exact position of document d within that specific list. Finally, $k = 60$ is the smoothing constant, preventing top-ranked documents in any single stream from dominating the fused score excessively.

3.3 Cross-Encoder Reranking

The top candidates from the fused retrieval pool are passed to a Cross-Encoder model (`BAAI/bge-reranker-v2-m3`). The Cross-Encoder computes a fine-grained relevance score

for each concatenated [Query + Document] pair. The system outputs the top-10 documents sorted by this reranking score.

4 Experimental Setup and Optimization

Our development phase focused on optimizing the pipeline over the answerable subset of the provided development data, aiming to maximize nDCG@5 under constrained computational resources.

4.1 Phase 1: Widening the Retrieval Funnel

In our initial configuration, the hybrid retriever fetched top-$K = 60$ candidates, passing the top 15 to the Cross-Encoder. This yielded a suboptimal nDCG@5 of 0.1744. A diagnostic analysis revealed a "funnel bottleneck": highly relevant documents were successfully retrieved by the embedding model but often ranked in the 70–100 range. By widening the initial retrieval pool to $K = 200$ and increasing the reranking threshold to 50, we successfully rescued these long-tail documents. This hyperparameter optimization led to a 14.2% performance jump, bringing the dev score to 0.1991.

4.2 Phase 2: Overcoming Topic Shifts

Qualitative evaluation of the Nano-LLM highlighted a critical failure mode: Topic Shifts. When users abruptly changed the subject, the base prompt forced the LLM to artificially inject outdated entities into the new query. We refined the prompt to explicitly detect pivots by adding strict behavioral rules (see Appendix A). Post-tuning analysis on 350 failure cases showed that only 12% of retrieval failures were caused by bad rewrites, demonstrating that a 1.5B model is highly capable of complex contextual resolution when guided correctly.

4.3 Phase 3: Weighted Fusion and Hybrid Tuning

Beyond widening the funnel, we implemented fine-grained weighting strategies to maximize precision:

- **Weighted Track Fusion:** When fusing the parallel retrieval streams via RRF, we assigned a higher multiplier to the *rewritten query track* compared to the *raw context track*. This minimized semantic "noise" introduced by long, rambling conversational histories.
- **Hybrid Alpha Tuning:** For highly technical domains (e.g., IBM Cloud), semantic dense embeddings often fail to match exact product codes or acronyms. We adjusted the hybrid alpha parameter, granting higher weight to the sparse (lexical) representations to ensure exact-keyword matching for technical terminology.

Corpus Domain	Count	nDCG@5
Govt	157	0.1759
ClapNQ	142	0.1603
IBM Cloud	131	0.1209
FIQA (Finance)	77	0.0994

Table 1: Performance breakdown by domain on the evaluation subset.

5 Results and Discussion

The system was evaluated on the blind test set provided by the SemEval organizers.

5.1 Performance and Resource Efficiency

Our official submission yielded an nDCG@5 score of 0.1744 (Rank 35/38). For context, the top-performing baseline achieved 0.4795 using the commercial ELSER engine combined with a massive 20B parameter LLM (GPT-OSS-20b) for query rewriting.

Our internal diagnostics revealed that out of 507 answerable queries, our pipeline successfully retrieved at least one relevant gold document within the top-5 candidates in 74% of the cases (373 queries). However, the overall nDCG@5 score was bottlenecked because these relevant documents frequently landed in ranks 3–5 rather than the top position, and in cases with multiple relevant documents, the system often failed to retrieve all of them within the top 5. To better understand this limitation, we expanded the evaluation window: our system achieves a Recall@10 of 0.2631 and an nDCG@10 of 0.1851. These metrics confirm that while the system is highly effective at surfacing an initial relevant document, it struggles to capture the complete set of relevant passages required for comprehensive multi-document queries, missing several relevant documents that are either ranked lower or unretrieved.

Our system deliberately prioritized extreme resource efficiency, utilizing a 1.5B parameter Nano-LLM. Operating with less than 8% of the parameters of the top baseline, our system still successfully

Query Type	Conversation History & Rewriting	System Behavior & Outcome
Answerable (Full Success)	**History:** [...omitted...] User: Does physical activity increase a child's attention span? AI: Yes, physical activity can increase... **Last Q:** How much physical activity a day is recommended for children? **Rewritten:** What is the daily recommended amount of physical activity for children?	**Outcome: Success.** The Nano-LLM correctly identified that the last question was largely self-contained. It perfectly isolated the core intent without getting distracted by the previous conversational noise regarding "attention spans" or "grades". Consequently, the hybrid retriever matched the exact semantic intent and ranked the golden Govt document at Position #1.
Answerable (Retrieval Gap)	**History:** [...omitted for brevity...] User: By the way, what is a secret? AI: A secret is any piece of data... **Last Q:** I heard the toolchain is not available in South America. **Rewritten:** Is the toolchain currently available in South America?	**Outcome: Failure.** The Nano-LLM perfectly converted a conversational statement into a well-formed standalone query. However, the retriever failed to surface the gold document because the text explicitly listed "Sao Paulo" rather than "South America". This exposes a semantic world-knowledge gap in the retrieval stage, despite a flawless rewrite.
Answerable (Rewrite Failure)	**History:** User: EV used car market AI: I do not have specific information... **Last Q:** I mean current EV's battery does not stand for a used car market...how do you think? **Rewritten:** I am looking for information on the current electric vehicle (EV) used car market. How do you perceive its potential compared to traditional gasoline vehicles?	**Outcome: Failure.** The Nano-LLM dropped the critical keyword "battery" and hallucinated a comparison to "gasoline vehicles". Consequently, the retriever fetched irrelevant general EV market articles, missing the gold FIQA forum posts discussing battery chemistry and longevity. This highlights the vulnerability of generative rewriting to hallucination in specialized domains.
Underspecified (Hallucinated Intent)	**History:** [...omitted...] User: Is Shylock a villain or a victim? AI: The interpretation of Shylock's character... **Last Q:** Do you believe this ocean currents play a crucial role in global climate regulation? **Rewritten:** Do you think ocean currents significantly impact global climate regulation?	**Outcome: Failure.** The user abruptly shifted topics from Shakespeare to "this ocean", creating an underspecified query. Instead of abstaining, the Nano-LLM over-corrected by removing the ambiguous demonstrative ("this") and generated a highly coherent, generic question. Consequently, the retriever fetched completely irrelevant documents, resulting in a severe false-positive penalty.

Table 2: Case studies from system logs illustrating a full success, a semantic retrieval gap, a rewriting hallucination in a specialized domain, and a critical vulnerability to underspecified queries due to generative over-correction.

navigated complex multi-turn dependencies. The drop from 0.1991 (Dev) to 0.1744 (Official Test) indicates a natural domain shift and highlights the difficulty Nano-LLMs face when generalizing to unseen conversational structures without massive parameter memorization.

5.2 Domain Discrepancy

An analysis across the four corpora (Table 1) reveals that the system excelled in the **Govt** domain (0.1759) but failed dramatically in the **FIQA** domain (0.0994).

The Government corpus features formal, structured language that aligns well with the general-purpose `BGE-m3` embeddings. In contrast, FIQA relies heavily on niche financial jargon and informal forum structures, proving that domain-agnostic dense embeddings are insufficient without domain-specific fine-tuning.

5.3 Qualitative Error Analysis: Answerable vs. Underspecified Queries

To better understand the discrepancy between our internal development score and the official test score, we conducted an analysis on the organizers' surprise "Underspecified" class (Rosenthal et al., 2026a). We note that these underspecified queries are used exclusively for diagnostic analysis and are not part of the questions used to compute the official leaderboard scores.

As demonstrated in Table 2, the generative query rewriter acts as a double-edged sword. For **Answerable** queries, the Nano-LLM successfully resolves ambiguous pronouns from the history, allowing the hybrid retriever to fetch the exact golden document.

Conversely, for **Underspecified** queries, the user's intent is inherently vague or the required information simply does not exist in the domain corpus. Instead of halting, the Nano-LLM rewriter "hallucinates" a highly coherent, plausible-sounding standalone query based on conversational

context. Because the rewritten query is grammatically perfect and semantically dense, the `BGE-m3` retriever successfully finds documents that sound similar but are entirely irrelevant (hard false-positives). This mechanism explains the severe penalty incurred on the blind test set: the better the LLM is at rewriting, the more confidently the system retrieves incorrect information when the underlying question is unanswerable.

6 Conclusion

We presented a multi-turn RAG pipeline demonstrating that a Nano-LLM (1.5B) can effectively rewrite complex conversational queries. While absolute performance trails larger 20B baselines, our system establishes a strong benchmark for local, resource-efficient CIS, overcoming the lexical BM25 baseline (nDCG@5 of 0.18) on answerable queries. Our error analysis highlights the vulnerability of generative rewriting to underspecified queries and out-of-domain vocabulary (FIQA). Future work will focus on integrating domain-specific fine-tuning and intent-classification to dynamically bypass the rewriter for unanswerable queries. Specifically, this abstention mechanism could be implemented by training a lightweight classifier to detect vague demonstratives or out-of-context pivots before the rewriting stage, routing such queries directly to a user-clarification prompt rather than the retrieval pipeline.

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.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624*.

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.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1933–1936.

A Prompt Template for Query Rewriting

System Prompt:
You are an expert Query Rewriter for a Retrieval-Augmented Generation system. Your task is to rewrite the last user question to be a fully self-contained, standalone query that can be used for semantic search. You must interpret the user's question based on the conversation history.

Rules:

1. **Entity Replacement:** Replace pronouns (it, this, he, she, they) with specific entities from the history.
2. **Reference Resolution:** Resolve ambiguous references (e.g., "the second one", "that tool").
3. **Context Inclusion:** If the user asks a follow-up question (e.g., "Why?", "How much?"), include the context.
4. **Output Constraint:** DO NOT answer the question. Output ONLY the rewritten query.
5. **Identify Topic Shifts:** If the current question introduces a completely new topic unrelated to the history, do not incorporate previous entities. Keep the new query standalone as is.
6. **Prioritize Intent:** If the last question contradicts the previous context or clearly pivots (e.g., "Actually, let's talk about..."), ignore the old context and focus on the new intent.

Examples:

History:
User: Tell me about the Qualified Invoice System.
AI: It is a system for consumption tax...
User: When did it start?

Rewritten: When did the Qualified Invoice System start?

History:
User: What is the difference between RAG and Fine-tuning?
AI: RAG retrieves data... Fine-tuning updates weights...
User: Which one is better for dynamic data?
Rewritten: Is RAG or Fine-tuning better for dynamic data?

History:
User: Tell me about IBM Cloud security features.
AI: It includes IAM, VPC security groups, and encryption...
User: Actually, I meant to ask about AWS S3 buckets.
Rewritten: What are the security features of AWS S3 buckets?