# UTRAG at SemEval-2026 Task 8:

**Ke Zhou*, Yi-Shan Lin***
Department of Linguistics, University of Tübingen
{ke.zhou, yi-shan.lin}@student.uni-tuebingen.de

## Abstract

This paper presents our system for the SemEval-2026 Task 8 MTRAGEval: Evaluating Multi-Turn RAG Conversations. The shared task focuses on answering context-dependent questions in multi-turn conversations, where identifying relevant dialogue history is crucial for effective retrieval. In this work, we propose a history-aware retrieval approach built on dense indexing. Given the last turn, our system selects relevant history questions by combining semantic similarity with recency, and rewrites the current query into a standalone form for retrieval. We additionally explore decomposing queries into multiple sub-questions to improve retrieval coverage (with effectiveness yet to be fully validated). Our approach aims to (affect). We achieve competitive results in the shared task:(if it performs well, we can write the records here). (but if not-The source code used in this paper is available at github.)

## 1 Introduction

Large Language Models (LLMs) are increasingly deployed as conversational assistants for information-seeking tasks. To improve factual grounding and reliability, Retrieval-Augmented Generation (RAG) has become a widely adopted paradigm, where external documents are retrieved and used as evidence during generation. While RAG has been extensively studied in single-turn question answering, recent work shows that multi-turn conversations introduce additional challenges that are not captured by single-turn benchmarks, such as non-standalone questions, reference to earlier turns, answerability, and changing information needs across a dialogue. For example, a user may first ask “Where does Doctor Strange get his powers from?” and later follow up with “How many films does he appear in?”, where the latter question is non-standalone and relies on entities introduced earlier in the dialogue.

The recently proposed MTRAG benchmark highlights these challenges by providing human-generated multi-turn conversations across multiple domains, and demonstrates that even state-of-the-art RAG systems struggle on later turns and context-dependent queries. To systematically evaluate these issues, the SemEval-2026 Task 8 (MTRAGEval) is organized with three subtasks focusing on retrieval, generation with reference passages, and full RAG evaluation.

In this paper, we present our methods for improving retrieval, generation, and the overall pipeline, as part of our participation in the SemEval Task 8 challenge.

The remainder of this paper is organized as follows. Section 2 reviews related work, introduces the MTRAG benchmark. Section 3 presents an overview of our system, covering our approaches for all subtasks. Section 4 reports the evaluation results. Finally, Section 5 concludes the paper.

## 2 MTRAG Benchmark/Related Work

To facilitate the evaluation of Retrieval-Augmented Generation in multi-turn conversational settings, Katsis et al. (2025) introduce MTRAG, a human-generated multi-turn RAG benchmark designed to reflect real-world information-seeking dialogues. MTRAG consists of 110 conversations with an average of 7.7 turns per conversation, leading to 842 tasks. Each task includes the full dialogue history up to the current turn, together with the last user question, enabling evaluation under context-dependent conditions.

In the MTRAG benchmark, each corpus is indexed using Elasticsearch with the ELSER retriever (Katsis et al., 2025). Documents are split into passages of 512 tokens, and the passages are used for retrieval and overall RAG pipeline. As

[*] : equal contribution.

Table 1: Statistics of the MTRAG benchmark, based on data from the official MTRAG repository (Katsis et al., 2025).

| Corpus | Domain | Documents | Passages |
|---|---|---|---|
| ClapNQ | Wikipedia | 4,293 | 183,408 |
| Cloud | Technical Documentation | 57,638 | 61,022 |
| FiQA | Finance | 7,661 | 49,607 |
| Govt | Government | 8,578 | 72,422 |

shown in Table 1, the four corpora cover different domains and vary in scale and granularity. At the task level, most instances are multi-turn follow-up questions, with summarization and explanation being the most common question types. Tasks are predominantly answerable, and the four corpora cover diverse domains with substantially different scales.

## 3 System Overview

Our system is designed for multi-turn Retrieval-Augmented Generation (RAG) and addresses Subtasks A–C of SemEval Task 8. Given the current user query and the dialogue history, we first perform history-aware query rewriting to obtain a standalone query for retrieval. We additionally explore query decomposition into multiple sub-queries for reranking as an auxiliary strategy. Finally, the selected evidence passages are fed into a LoRA-adapted generator for response generation.

### 3.1 Corpus and Query Dense Embedding

We consider three official baseline retrievers provided by the task, namely BM25, BGE-base, and ELSER. In our system, we adopt dense indexing as the primary retrieval approach, where both corpus passages and queries are embedded into the same vector space and retrieved via nearest-neighbor search.

### 3.2 History-Aware Query Rewriting (Subtask A)

Multi-turn user queries are often non-standalone and depend on entities or constraints introduced in earlier turns. To mitigate this issue, we rewrite the current query into a standalone form by incorporating a compact set of relevant context from the dialogue history.

To select relevant context for query rewriting, we compute dense embeddings for the current query and all candidate history questions. Cosine similarity is used to measure semantic relatedness between the current query and each history turn. We always retain the most recent history question to preserve local context, and select additional history questions based on similarity ranking.

The selected history questions are then combined with the current query and rewritten into a standalone question using a constrained language model prompt. The rewritten query is finally used as the input to the retriever for document and passage retrieval.

#### 3.2.1 Pronoun Resolution

During query rewriting, we use dense embeddings to identify the most relevant dialogue history in order to transform the current query into a standalone form. However, pronouns in conversational queries may negatively affect embedding-based similarity, as their referents are often ambiguous across turns. To address this issue, we perform pronoun resolution on dialogue history queries before history selection.

Specifically, we apply a history-aware pronoun rewriting strategy that replaces pronouns with explicit entity mentions derived from relevant earlier turns. This process yields more explicit queries, which in turn facilitates more accurate retrieval.

### 3.3 Sub-Query Decomposition for Reranking (Subtask A)

To better capture fine-grained information needs, we further decompose the rewritten query into a small set of sub-queries. Each sub-query retrieves a candidate set of passages, and the union of candidates is then reranked to produce the final evidence set. This decomposition is used as an auxiliary strategy and is intended to improve evidence coverage and robustness for context-dependent queries.

### 3.4 LoRA-Adapted Generator (Subtask B)

For generation, we use a decoder-based model and apply parameter-efficient fine-tuning with LoRA. The generator is trained via supervised fine-tuning

(SFT) on the provided reference data, which is split into training and development sets. At inference time, the generator conditions on the current user query together with one or two previous dialogue turns as conversational context, as well as the provided reference passages, to produce the final response.

### 3.5 End-to-End RAG Pipeline (Subtask C)

The full pipeline integrates rewriting, retrieval, reranking, and generation.

## 4 Experimental Setup

### 4.1 Data and Evaluation Methodology

### 4.2 Model and Implementation

## 5 Results

## 6 Conclusion

## References

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. *Preprint*, arXiv:2501.03468.