IITKanBDone at SemEval-2026 Task 8: ELSER-Based Sparse Retrieval with Mistral-7B for Multi-Turn RAG Evaluation

Garima Gupta
Indian Institute of Technology Kanpur
Department of Electrical Engineering
gupta.garima2@gmail.com

Soumendra Nath Ray
Indian Institute of Technology Kanpur
Department of Electrical Engineering
soumendra.nath.ray@gmail.com

Abstract

This paper describes our system for the MT-RAG (Multi-Turn Retrieval-Augmented Generation) shared task (Rosenthal et al., 2026b), which addresses the challenge of multi-turn conversational question answering using retrieval-augmented generation. We participated in three sub-tasks of Task 8: Task A (retrieval), Task B (generation with reference passages), and Task C (end-to-end RAG). For Task A, we evaluated multiple retrieval approaches including BM25, BGE, and hybrid methods, achieving best performance with ELSER (Elastic Learned Sparse EncodeR) with nDCG@5 of 0.4018 (Rank 24/38). For Task B, we employed the Mistral-7B-Instruct-v0.2 model via HuggingFace for response generation using gold reference passages, achieving a harmonic mean score of 0.6976 (Rank 13/26). For Task C, we combined ELSER retrieval with Mistral-7B generation, using top-5 retrieved passages as context, achieving a score of 0.4289 (Rank 23/29). Our system demonstrates the effectiveness of learned sparse retrieval methods and instruction-tuned models for multi-turn conversational RAG scenarios.

1 Introduction

The MT-RAG (Multi-Turn Retrieval-Augmented Generation) shared task focuses on evaluating RAG systems in multi-turn conversational settings over UN Documents, addressing a critical challenge in building interactive question-answering systems (Rosenthal et al., 2026b).

Main Strategy: Our system adopts a pipeline approach combining learned sparse retrieval (ELSER) with instruction-tuned language models (Mistral-7B-Instruct-v0.2). For retrieval (Task A), we systematically evaluated BM25, BGE embeddings, and ELSER, selecting ELSER as our final approach. For generation (Tasks B and C), we use the Mistral-7B-Instruct-v0.2 model with top-5 reference/retrieved passages as context.

Key Findings: Our experiments demonstrated that ELSER outperformed both traditional sparse (BM25) and dense (BGE) retrieval methods (Fadnis et al., 2025). Our Task B generation achieved competitive results (Rank 13/26), outperforming the gpt-oss-120b baseline despite using a significantly smaller 7B parameter model. The modular pipeline design allowed independent optimization of retrieval and generation components.

Code Availability: Our implementation is available at https://github.com/guptagarima2/mtrag.

The MT-RAG shared task is described in detail in the task description paper (Rosenthal et al., 2026b).

2 Background

The SemEval-2026 Task 8 focuses on Multi-Turn Retrieval-Augmented Generation (MT-RAG) over UN Documents (Rosenthal et al., 2026a; Katsis et al., 2025). The task evaluates systems' ability to retrieve relevant information and generate accurate responses across multiple conversation turns using UN documentation as the knowledge base (Rosenthal et al., 2026b).

Retrieval-Augmented Generation (RAG) has emerged as a powerful approach for enhancing large language models by grounding their responses in retrieved factual information (Lewis et al., 2020). Traditional RAG systems typically handle single-turn queries, but multi-turn conversations present additional challenges including context tracking, query reformulation, and maintaining coherence across dialogue turns.

The task comprises three subtasks: **Task A** (Retrieval System) – identifying relevant passages from UN documents, evaluated using nDCG@5; **Task B** (Answer Generation) – generating answers using gold reference passages, evaluated using the harmonic mean of R_Bagg, R_LF, and R_Bllm;

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1063–1067

and **Task C** (End-to-End RAG) – complete pipeline from query to answer generation (same evaluation as Task B).

The evaluation dataset is based on UN documentation spanning multiple topics including sustainable development, human rights, and international cooperation. The task also includes an underspecified answerability class as a surprise element, where questions cannot be fully answered from the provided context (Rosenthal et al., 2026b).

3 System Overview

Our system follows a modular pipeline approach, addressing each task independently while maintaining consistency in methodology.

3.1 Task A: Retrieval System

We experimented with multiple retrieval approaches: We evaluated four retrieval approaches: (1) **BM25** – traditional sparse retrieval using BM25 algorithm; (2) **BGE** – BGE-base-1.5 embeddings (Wolf et al., 2020) for semantic similarity; (3) **Hybrid** – a combination of BM25 and BGE scores; and (4) **ELSER** (our final choice) – Elastic Learned Sparse EncodeR v2, implemented via Elasticsearch (Elastic, 2023).

Implementation Details: The system used Elasticsearch 8.10 with the ELSER model (`.elser_model_2`), executing text expansion queries with the `model_text` parameter. We retrieved the top-10 documents per query. Naive concatenation with limited context performed similarly so for simplicity we used only the last user turn from the multi-turn conversation as the query. Documents were indexed on a per-corpus basis with document ID and text fields (Elastic, 2023).

3.2 Task B: Generation with Reference Passages

For Task B, given gold reference passages, we generate responses using: We used the Mistral-7B-Instruct-v0.2 model (Jiang et al., 2023) via HuggingFace Transformers (Wolf et al., 2020). Note that Mixtral-8x7B was not allowed as it was used for dataset creation. The context consisted of gold reference passages provided in the task. The prompt included the full conversation history (all previous turns), reference passages formatted with IDs, and instructions emphasizing faithfulness and avoiding hallucination.

For generation parameters we used a temperature of 0.3 for more deterministic outputs, a maximum of 512 tokens, and enabled sampling.

3.3 Task C: End-to-End RAG Pipeline

Task C combines retrieval and generation. Our pipeline combined ELSER-based retrieval from Task A (top-5 passages) with the same Mistral-7B-Instruct-v0.2 model as Task B. The pipeline followed five steps: (1) extract the last user query from the conversation for use in concatenated prompt in step 4; (2) retrieve top-K passages using ELSER (Elastic, 2023); (3) enrich contexts with passage text from the corpus; (4) build a RAG prompt combining conversation history with retrieved contexts; and (5) generate a response using the LLM (Jiang et al., 2023).

4 Experimental Setup

Data: We used the official MT-RAG benchmark datasets provided by the task organizers (Rosenthal et al., 2026b). The data consists of multi-turn conversations across multiple UN document collections/domains.

External Tools and Libraries: The implementation used Elasticsearch 8.10 (Elastic, 2023) for ELSER-based retrieval, HuggingFace Transformers 4.x (Wolf et al., 2020) for LLM inference, Python 3.8 as the core implementation language, and PyTorch as the backend for transformer models.

Hyperparameters: For retrieval, we retrieved the top-K = 10 documents for Task A and top-K = 5 passages for Tasks B & C. For generation, we used a temperature of 0.3 and a maximum of 512 tokens. The Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) model was used without fine-tuning in a zero-shot inference setting.

Evaluation Metrics: Task A was evaluated using nDCG@5 (normalized Discounted Cumulative Gain). Tasks B & C were evaluated using the harmonic mean of R_Bagg (ROUGE-based aggregate), R_LF (ROUGE-L F1), and R_Bllm (LLM-as-judge score) (Rosenthal et al., 2026b).

5 Results

5.1 Official Results

Table 1 presents our official results on the MT-RAG shared task leaderboard.

Task	Metric	Our Score	Rank	Top Score
Task A	nDCG@5	0.4018	24/38	0.5776
Task B	Harm. Mean	0.6976	13/26	0.7827
Task C	Harm. Mean	0.4289	23/29	0.5861

Table 1: Official Leaderboard Results

5.2 Task B Detailed Metrics

Table 2 presents our detailed metrics on Task B where our results outperformed the baseline.

Metric	Score
R_Bagg (ROUGE-based aggregate)	0.5753
R_LF (ROUGE-L F1)	0.7865
R_Bllm (LLM-as-judge)	0.7746
Harmonic Mean	0.6976

Table 2: Task B Detailed Metrics

5.3 Task C Detailed Metrics

Table 3 presents our detailed results for Task C (End-to-End RAG), combining ELSER-based retrieval with Mistral-7B-Instruct-v0.2 generation.

Metric	Score
R_Bagg (ROUGE-based aggregate)	0.3100
R_LF (ROUGE-L F1)	0.6082
R_Bllm (LLM-as-judge)	0.4707
Harmonic Mean	0.4289

Table 3: Task C Detailed Metrics

5.4 Retrieval Method Comparison

During development, we evaluated multiple retrieval approaches on the baseline dataset (Table 4).

Method	R@10	nDCG@10	% of Baseline
BM25	0.263	0.205	97%
BGE (dense)	0.385	0.308	101%
ELSER (sparse)	0.544	0.450	94%

Table 4: Retrieval Method Comparison (Development)

Our ELSER results achieved 94% of the paper's reported performance, while significantly outperforming both lexical (BM25) and dense (BGE) retrieval methods in our setup, validating our choice for the final submission.

5.5 Analysis

Task B Performance: Our Task B result (0.6976) exceeded the gpt-oss-120b baseline (0.639), demonstrating that smaller instruction-tuned models like Mistral-7B can achieve competitive generation quality when provided with gold reference passages. The high R_LF score (0.7865) indicates good lexical overlap with reference answers.

Task A vs Task C Gap: The significant gap between Task B (0.6976) and Task C (0.4289) scores highlights the impact of retrieval quality on end-to-end RAG performance. Our retrieval nDCG@5 (0.4018) using naive concatenation with prompt engineering was below the top baseline using query rewriting (0.4795), suggesting that query reformulation for multi-turn conversations is a key area for improvement.

Model Size Trade-off: Despite using a smaller model (Mistral-7B, 7B parameters) compared to larger alternatives (GPT-OSS-120b, Qwen-30b), our system achieved competitive results, particularly on Task B. This demonstrates the viability of efficient, smaller models for RAG applications.

5.6 Error Analysis

Our error analysis identified several common failure patterns across the tasks.

Retrieval Errors (Task A): The primary retrieval failure mode was query-document vocabulary mismatch in multi-turn contexts, where the last user turn lacked sufficient context for disambiguation. Single-turn query extraction caused ELSER to miss important contextual coreferences from conversation history. Naive concatenation of the last two dialogues had a marginal impact. Additionally, ELSER's learned sparse representations occasionally failed to capture domain-specific UN terminology.

Generation Errors (Tasks B and C): The generation module exhibited hallucination in responses despite clearly provided passages. We also observed difficulty in synthesizing information across multiple passages for complex multi-hop questions. Query re-writing had to choose a limited context, so we kept it out for simpler analysis. We focused on prompt engineering, but handling of under-specified questions where full answers were not available in context posed challenges, particularly for the surprise underspecified answerability class (Rosenthal et al., 2026a). Furthermore, context window limitations with temperature=0.3 (Task B) or 0.7 (Task C) occasionally affected coherence in longer responses.

5.6.1 Qualitative Case Studies

To complement the aggregate error metrics, we include two concrete failure cases from the evaluation set.

Case 1 (Task ID: `016cae9db564f372edbaf919e0a581b0<::>7`) illustrates multi-turn coreference failure. The final user query ("is this song famous in Asia as well?") refers to a previously discussed U2 song, but retrieval returns a passage about Anggun ("La Neige au Sahara"). Conditioned on this off-target evidence, the Task C generator confidently answers about Anggun, indicating an entity-switch hallucination caused by retrieval drift. In contrast, Task B for the same item responds conservatively ("I don't have enough information"), consistent with missing/empty reference context.

Case 2 (Task ID: `24ecd67f930a4927f40ed7dae21ca600<::>2`) illustrates under-specified follow-up failure. The query ("What sporting events are hosted in the country's stadium?") lacks a resolved country reference from conversation history; retrieval surfaces noisy context (e.g., San Jose SAP Center snippet), while Task C synthesizes unsupported claims across unrelated entities (e.g., Wales, Mexico City). This reflects a retrieval-disambiguation failure followed by generation overgeneralization.

Result: Together, these examples concretely demonstrate the two dominant patterns observed in our error statistics: context-resolution failures in retrieval and unsupported synthesis in generation.

6 Conclusion

This paper presented our system for the shared task SemEval-2026 MT-RAG (Task 8), focusing on the generation augmented by multi-turn retrieval over UN documents (Rosenthal et al., 2026b). Our approach combines ELSER-based retrieval with the Mistral-7B-Instruct-v0.2 model for response generation.

Key Results: Our system achieved nDCG@5 of 0.4018 in Task A (Retrieval, Rank 24/38), a harmonic mean of 0.6976 in Task B (Generation, Rank 13/26), exceeding the gpt-oss-120b baseline, and a harmonic mean of 0.4289 in Task C (End-to-End RAG, Rank 23/29).

Key Contributions: This work systematically evaluated retrieval methods (BM25, BGE, ELSER), demonstrating ELSER's effectiveness for this task. We implemented a modular pipeline architecture that allows independent optimization of the retrieval and generation components. Furthermore, we demonstrated that smaller models (Mistral-7B) can achieve competitive generation performance with proper prompt engineering (Jiang et al., 2023).

Limitations: Our retrieval approach uses only the last user query in prompt with naive concatenation, which still misses valuable context from conversation history. We did not implement a query rewriting mechanism reliably, which the top baseline effectively employed. The system is limited to zero-shot inference without domain-specific fine-tuning on UN documents, and lacks an explicit answerability detection module for handling under-specified questions (Rosenthal et al., 2026a).

Future Work: Future work will optimize query rewriting for conversation-aware query reformulation using LLMs (the top baseline achieved 0.4795 vs our 0.4018). Other promising directions include combining ELSER with dense embeddings for complementary retrieval signals (Fadnis et al., 2025), integrating conversation history into the retrieval query, adding a cross-encoder re-ranker to improve retrieval precision, fine-tuning retrieval and generation models on UN documentation, and developing strategies to detect and appropriately respond to underspecified questions (Rosenthal et al., 2026a)

Acknowledgments

We would like to thank our mentor at IIT Kanpur, Professor Dr Ashutosh Modi for his guidance and support throughout this work. We also acknowledge the SemEval-2026 Task 8 organizers for providing the MT-RAG benchmark and evaluation framework.

References

Elastic. 2023. Elastic learned sparse encoder (elser). Accessed: 2024.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux,

Arthur Mensch, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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 Bouza, Alexandros Konstantinidis, Zhengzhong Ren, Kevin Murray, Kevin Murphy, Walid Soubra, Jeremy Devers, Younes Belkada, Aman Acharya, and Santanu Brahma. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R
'emi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.