# NITATREC at TREC RAG 2025: A Report on Exploring Sparse, Dense, and Hybrid Retrieval for Retrieval-Augmented Generation

**Aparajita Sinha**
National Institute of Technology
Agartala, Tripura, India
asinha.22odcsw002@phd.nita.ac.in

**Kunal Chakma**
National Institute of Technology
Agartala, Tripura, India
kchakma.cse@faculty.nita.ac.in

## Abstract

This paper describes our participation in the TREC RAG 2025 shared task, which investigates Retrieval-Augmented methods for addressing complex information needs using the MS MARCO v2.1 document and segment collections. We submitted systems to all four subtasks: Retrieval (R), Augmented Generation (AG), Retrieval-Augmented Generation (RAG), and Relevance Judgment. For the retrieval task, we explored three approaches: a lexical BM25 baseline, a dense retrieval model based on DPR embeddings, and a hybrid pipeline combining sparse and dense retrieval with cross-encoder reranking. For the generation tasks, we employed an instruction-tuned language model to produce evidence-grounded responses with citations. Experimental results show that the hybrid retrieval system achieves the best performance, obtaining a MAP of 0.1037, an nDCG@30 of 0.527, and a Recall@100 of 0.158. In the generation tasks, the RAG system achieved a Strict Vital Score of 0.19 and a Weighted Precision/Recall of 0.472. In contrast, the AG submission had a weighted precision/recall of 0.481. These results highlight the importance of combining lexical and semantic retrieval signals to improve Retrieval-Augmented Generation.

## 1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a highly effective framework for addressing complex information-seeking challenges that require both Retrieval and Natural Language Generation (Lewis et al., 2020; Guu et al., 2020). Traditional information retrieval (IR) systems are capable of locating significant documents or passages within extensive collections; however, they typically provide users with sorted lists that they must manually compile. Conversely, Large Language Models (LLMs) have the capacity to provide coherent and fluent responses; however, they may also fabricate or provide inaccurate information in the absence of other sources of information. RAG systems resolve these issues by integrating generative models with retrieval methods. This allows for responses to be derived from evidence that has been retrieved (Lewis et al., 2020; Gao et al., 2023). Relevant text segments are obtained from a corpus by an IR component of these systems, and an LLM generates responses based on this information. The TREC RAG 2025 shared task[1] provides a comprehensive assessment methodology for the evaluation of retrieval-augmented generation systems in authentic information retrieval contexts. The MS MARCO v2.1 segment collection (Nguyen et al., 2016) is used for this track. It has millions of small text segments taken from web documents using a sliding window segmentation approach. The subjects in this track are non-factoid, multi-sentence questions that show complicated information needs and often need to combine evidence from several parts (Thakur et al., 2025). The shared job breaks down into multiple subtasks, each of which corresponds to a distinct step in the Retrieval Generation pipeline. The Retrieval (R) subtask is all about getting the right parts of the corpus for each query. The Augmented Generation (AG) subtask separates the generation part by making systems come up with replies only from evidence segments that the organizer gives. The Retrieval-Augmented Generation (RAG) subtask combines Retrieval and Generation. It asks computers to find relevant parts and make grounded responses with citations to back them up. The Relevance Judgment subtask also lets participants add relevance annotations to pairings of topics and segments that will be used in the evaluation process.

We took part in all four subtasks of the TREC RAG 2025 track and built many systems that

[1] https://trec-rag.github.io/annoucements/2025-track-guidelines/

looked into different ways to retrieve and generate information. For the Retrieval subtask, we created three retrieval paradigms: a lexical Retrieval baseline that uses the BM25 ranking function (Robertson et al., 2009), a Dense retrieval model that uses MiniLM embeddings (Wang et al., 2020) and FAISS indexing (Douze et al., 2025), and a Hybrid retrieval system that uses both sparse and dense retrieval signals and then cross-encoder reranking.

The BM25 ranking function uses word frequency, inverse document frequency, and document length normalization to figure out how relevant a document or segment $d$ is to a query $q$. The scoring function is defined as follows:

$$\text{BM25}(q,d) = \sum_{t \in q} IDF(t) \times \frac{f(t,d)(k_1+1)}{f(t,d) + k_1\left(1 - b + b\frac{|d|}{avgdl}\right)} \quad (1)$$

where $f(t,d)$ indicates the frequency of the term $t$ within document $d$, $|d|$ denotes the document's length, $avgdl$ signifies the mean length of all documents in the corpus, and $k_1$ and $b$ are parameters that govern the scaling of term frequency and the normalization of document length. We constructed a pipeline for the Full RAG subtask, incorporating retrieval, evidence formatting, prompt construction, and answer generation, leveraging an instruction-tuned language model. The Falcon-7B (Aridoss et al., 2024) instruction-tuned model is utilized to pinpoint the top-$k$ pertinent segments, which are subsequently structured into queries. The system then produces grounded responses that incorporate citation references. We utilized the evidence segments provided by the organizer to generate structured JSONL outputs and replies for the Augmented Generation subtask, which were consistent with the text generated with citations. For the Relevance Judgment subtask, we implemented an automated relevance labeling pipeline. This pipeline employs a cross-encoder model to assign relevance scores to respective pairings of topics and segments. Numerous technical challenges emerge when utilizing the MS MARCO v2.1 segment collection. The preprocessing, embedding generation, and indexing pipelines must be engineered to handle a substantial data volume, given that the dataset comprises over 100 million segments. Consequently, the ability to process data expeditiously and leverage GPUs for computational tasks is essential for the large-scale generation of dense representations. The dataset's segmented format also complicates the process of combining and rating evidence. In terms of generation, the quality and foundation of the responses provided are significantly influenced by the following: rapid construction, evidence formatting, and citation alignment. The official results from the TREC RAG 2025 track evaluation are now available. This allows us to assess how well our systems performed on each individual subtask. In this report, we outline the design and implementation specifics of our Retrieval, Generation, and Relevance judgment pipelines. Furthermore, we share the evaluation outcomes for our submissions.

The main contributions of this work are summarized as follows:

- We evaluate and implement three retrieval paradigms for the TREC RAG Retrieval task: sparse BM25 retrieval, dense embedding-based retrieval, and hybrid retrieval with cross-encoder reranking.

- We create a Retrieval-Augmented Generation pipeline that employs an instruction-tuned language model to incorporate evidence retrieval, prompt construction, and citation-aware answer generation.

- We develop a Retrieval-Augmented Generation pipeline that integrates evidence retrieval, prompt formulation, and citation-aware answer production through the use of an instruction-tuned language model.

The succeeding sections of this work are organized as follows. The dataset is defined in Section 2 and a summary of the duties outlined in the shared track is provided. The system configurations and submitted trials for each subtask are detailed in Section 3. The evaluation findings are presented in Section 4, which also evaluates the system's performance in the Retrieval, Augmented Generation, and Full RAG tasks. Ultimately, Section 5 concludes the report and identifies potential areas for future research.

## 2 Methodology

### 2.1 Dataset Description

The TREC RAG 2025 shared task's official dataset is the **MS MARCO v2.1** segment collection, which serves as the foundation for all investigations in this work. The dataset is comprised of approximately 10.9 million web documents and 113.5 million text segments that were extracted from these documents using a sliding-window segmentation process. Each entry in the dataset is distributed as a gzipped JSONL record that contains metadata and identifiers, such as `docid`, `url`, `title`, `headings`, and `textual content`. The segmented corpus is produced by employing a sliding window of ten sentences with a stride of five sentences to the original documents. As a result, segments are typically composed of 500 to 1000 characters. These segment identifiers are the primary retrieval elements in all subtasks of the shared task. The official topic file contains **301 topics**. Each topic is identified by a `topic_id` and a concise narrative that expresses the information requirement.

### 2.2 Tasks Overview

The TREC RAG 2025 shared task is comprised of four subtasks: **Retrieval (R)**, **Augmented Generation (AG)**, **Retrieval-Augmented Generation (RAG)**, and **Relevance Judgment**. The official identifiers provided in the MS MARCO v2.1 segment collection are utilized by each task to guarantee the consistent evaluation and comparability of results. Participants are furnished with a topics file and the segment collection for the *Retrieval (R)* subtask. In the standard TREC run format, systems are required to return a ranked list of segments for each query. This list should include the following: `topic_id`, `Q0`, `segment_id`, `rank`, `score`, and `run_id`. The topic file and a ranked list of evidence segments generated by the coordinators are provided to participants for the *Augmented Generation (AG)* subtask. Responses must be generated by systems that are based on the evidence segments that have been provided, without the need to retrieve any additional documents. In the *Retrieval-Augmented Generation (RAG)* subtask, participants are responsible for both retrieval and generation. Systems generate grounded answers that are supported by the retrieved evidence by retrieving pertinent segments from the MS MARCO v2.1 collection. In the *Relevance Judgment* subtask, participants assign relevance labels to topic segment pairs in accordance with the annotation scheme outlined in the task guidelines.

### 2.3 System Pipeline

Our system integrates Sparse Retrieval, Dense Retrieval, Hybrid Ranking, and Retrieval-Augmented Generation components. Table 1 summarizes the main components used in the system pipeline, and Figure 1 illustrates the overall workflow.

Table 1: Overview of the retrieval and generation components used in our system pipeline.

| Component | Method / Model |
|---|---|
| Sparse Retrieval | BM25 (Anserini) (Robertson et al., 2009) |
| Dense Retrieval | MiniLM embeddings with FAISS indexing (Douze et al., 2025) |
| Fusion Strategy | Hybrid retrieval (Sparse + Dense) |
| Reranking | Cross-Encoder (MiniLM) (Reimers and Gurevych, 2019) |
| Generation | Instruction-tuned LLM with citation prompting |

The pipeline begins with query preprocessing and retrieval of candidate segments using a BM25 lexical retrieval model implemented through the Anserini framework. In parallel, dense vector representations for queries and document segments are generated using a MiniLM encoder.

### 2.4 Dense Retrieval

In dense retrieval, both queries and document segments are mapped into a shared embedding space using a neural encoder. Let $q$ denote the query embedding vector and $d$ denote the embedding vector of a document segment. The similarity between the query and document embeddings is computed using cosine similarity:

$$sim(q, d) = \frac{q \cdot d}{||q|| \, ||d||} \tag{2}$$

where $q \cdot d$ denotes the dot product between the query and document vectors, and $||q||$ and $||d||$ represent the corresponding vector norms. Higher similarity scores indicate greater semantic relevance between queries and candidate segments. The resulting embeddings are indexed using the FAISS library to enable efficient similarity search over the MS MARCO v2.1 segment collection.

### 2.5 Hybrid Retrieval and Reranking

To combine the advantages of lexical and semantic retrieval, we adopt a Hybrid retrieval strategy that merges candidate segments retrieved by BM25 and Dense embedding similarity. The merged candidate set is subsequently reranked using a cross-encoder model that jointly processes the query and candidate segment.

Table 2: Summary of our submitted retrieval runs.

| Run ID | Description |
|---|---|
| bm25_NITA_JH | BM25 lexical retrieval baseline |
| NITA_R_DPR | Dense retrieval using DPR embeddings |
| NITA_R_JH_HY | Hybrid retrieval with cross-encoder reranking |

The cross-encoder scoring function is defined as:

$$score(q, d) = CE([q; d]) \quad (3)$$

where $CE(\cdot)$ represents the cross-encoder model and $[q; d]$ denotes the concatenation of the query and document segment as a single input sequence. The cross-encoder outputs a relevance score that is used to reorder candidate segments.

### 2.6 Retrieval-Augmented Generation

The top-$k$ ranked segments generated by the hybrid retrieval model are utilized as supporting evidence for the generation of answers in the Full RAG task. An instruction-tuned language model is provided with these segments, which are formatted into structured prompts. The model generates responses that include citation references to the segments of supporting evidence. Ensuring that each statement can be traced back to the referred evidence passages, the generated answers are constrained to remain grounded in the retrieved segments. This design facilitates the transparent assessment of the quality of answer generation and the efficacy of retrieval. The official MS MARCO v2.1 segment identifiers are maintained in all subtasks to guarantee compatibility with the TREC RAG 2025 evaluation framework.

## 3 Submitted Tasks and System Runs

We submitted runs to all four subtasks defined in the TREC RAG 2025 shared task. Each subtask targets a different component of the retrieval generation pipeline, and our submissions explore different retrieval, generation, and relevance assessment strategies.

### 3.1 Retrieval (R) Only Submissions

**bm25_NITA_JH**. This iteration serves as a lexical baseline for the Retrieval (R) subtask. The BM25 (Robertson et al., 2009) ranking function, which is implemented through the Pyserini (Lin et al., 2021) framework, is utilized by the system to retrieve segments from the MS MARCO v2.1 segment collection. Queries are executed without any expansion or additional normalization. The system retrieves the top 100 segments for each topic, ranked by BM25 score. The process is entirely automated and employs only lexical matching, without the use of neural models.

**NITA_R_DPR**. The Dense Passage Retrieval (DPR) framework is employed to incorporate semantic retrieval in this run (Karpukhin et al., 2020). The DPR question encoder is employed to generate query embeddings, while the corresponding context encoder is employed to encode document segments. Segments that are semantically pertinent are identified by calculating the cosine similarity between embeddings. The final ranking is generated by reranking the best candidates using a cross-encoder model.

**NITA_R_JH_HY**. This experiment employs a hybrid multi-stage retrieval pipeline that integrates dense semantic retrieval with sparse lexical retrieval. BM25 is initially employed to retrieve a substantial candidate population of segments. Concurrently, dense retrieval based on DPR identifies segments that are semantically similar. The relevance score is computed by jointly encoding the query and segment in the `cross-encoder/ms-marco-MiniLM-L-12-v2` model (Reimers and Gurevych, 2019). Consequently, these candidate sets are merged and reranked. The final rated list is comprised of the top 100 segments following the reranking process.

### 3.2 Augmented Generation (AG) Submission

**NITA_AG_JH**. The answer generation process in the Augmented Generation (AG) subtask is entirely dependent on the evidence segments provided by the coordinators. This submission pertains to this subtask. The instruction-tuned language model `tiiuae/falcon-7b-instruct` (Almazrouei et al., 2023) is implemented by the system. The topic narrative is combined with the evidence segments and their identifiers to create a prompt for each topic. In order to guarantee that responses can be replicated, the model implements deterministic decoding parameters. The output is re-examined to identify and verify citation markers that correspond to the segment identifiers that were supplied. The final response is inserted into the AG Format-2 JSONL structure as specified by the assignment criteria. Duplicate or invalid citations are eliminated.

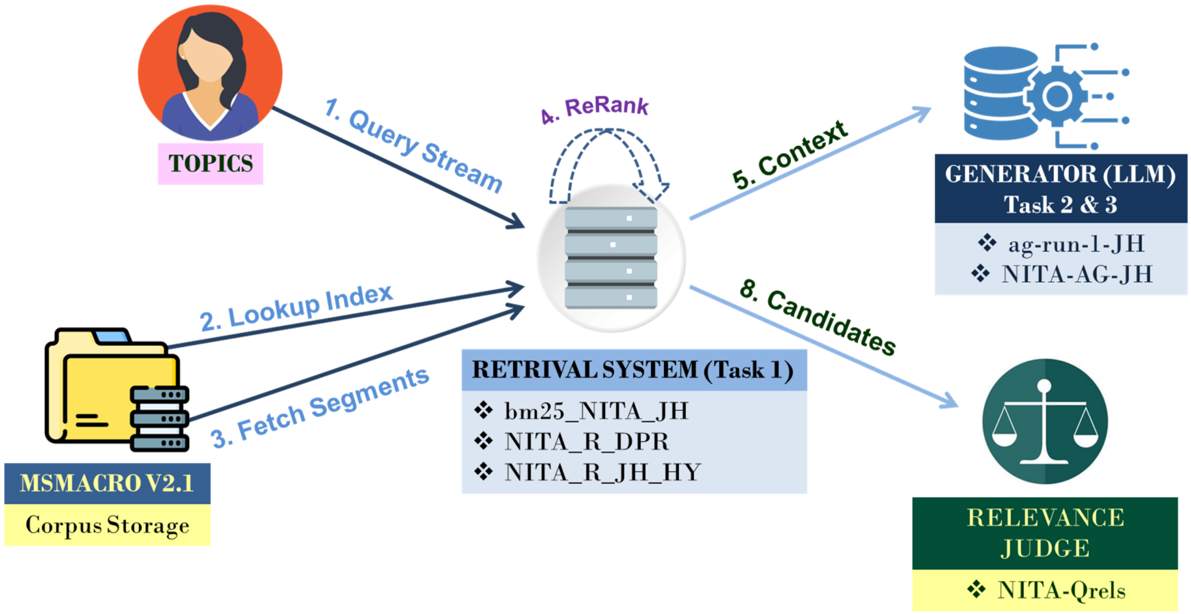

Figure 1: Overview of the methodology used in our TREC RAG 2025 submissions.

### 3.3 Retrieval-Augmented Generation (RAG) Submission

**ag-run-1-JH**. This submission pertains to the entire Retrieval-Augmented Generation (RAG) subtask, which entails the generation of answers and the acquisition of information. For each topic, a precomputed retrieval operation is conducted over the MS MARCO v2.1 segment collection to acquire a candidate set of segments. The best 20 segments are selected as proof passages and incorporated into a structured prompt. We employ the `tiiuae/falcon-7b-instruct` (Penedo et al., 2023) model with deterministic decoding parameters to generate responses. The citation markers that are linked to the supporting evidence segments are verified through post-processing, after the responses that were generated are divided into sentences. The final product is a JSONL file that contains the topic ID, prompt, evidence segment IDs, and citation-grounded replies.

### 3.4 Relevance Judgment Submission

**NITA-Qrels**. For the Relevance Judgment subtask, we conducted automated relevance assessments for topic-segment pairings. Candidate segments were selected from the top outcomes of a baseline dense retrieval run. Each query-segment pair was compared using the `BAAI/bge-reranker-large`[2] model. It assigns a score based on the degree of similarity between the topic narrative and the segment text in terms of their meaning. The similarity scores were assigned to a five-level relevance scale, with 0 representing "not relevant" and 4 representing "very relevant," based on predetermined thresholds. The labels that were produced were composed in the conventional TREC QREL format. They possessed a confidence score, a run identifier, a relevance designation, a segment identifier, and a subject identification.

## 4 Results

The official evaluation results for our submitted trials in the TREC RAG 2025 shared task are presented in this section. The results encompass the Retrieval (R), Augmented Generation (AG), Retrieval-Augmented Generation (RAG), and Relevance Judgment subtasks. The evaluation metrics are consistent with the standard TREC evaluation framework and assess the quality of generated responses as well as the effectiveness of retrieval.

### 4.1 Retrieval Task Results

The Retrieval (R) subtask assesses the efficiency with which a system can identify pertinent segments from the MS MARCO v2.1 segment collection. We submitted three retrieval trials that represented distinct retrieval paradigms: a lexical baseline that utilized BM25, a dense retrieval system that was based on DPR embeddings, and a hybrid retrieval approach that combined sparse and

[2] https://huggingface.co/BAAI/bge-reranker-large/discussions/11

Table 3: Retrieval performance of the BM25 baseline run.

| Metric | Score (bm25_NITA_JH) |
|---|---|
| MAP | 0.0571 |
| MRR | 0.6641 |
| nDCG | 0.1615 |
| nDCG@30 | 0.3966 |
| P@10 | 0.4909 |
| Recall@100 | 0.1120 |

Table 4: Retrieval performance of the dense retrieval run.

| Metric | Score (NITA_R_DPR) |
|---|---|
| MAP | 0.0874 |
| MRR | 0.8235 |
| R-Precision | 0.1333 |
| nDCG | 0.1947 |
| nDCG@30 | 0.4864 |
| nDCG@100 | 0.3586 |
| P@10 | 0.6409 |
| Recall@100 | 0.1333 |

dense retrieval with neural reranking.

#### 4.1.1 BM25 Baseline Retrieval

Table 3 reports the performance of the BM25 lexical baseline. While BM25 is effective for capturing exact term matches between queries and segments, it is limited in its ability to identify semantically related passages.

#### 4.1.2 Dense Retrieval using DPR

As shown in Table 4, the dense retrieval approach improves performance over the lexical baseline across most ranking metrics. By representing queries and segments using semantic embeddings, DPR is able to capture conceptual similarity even when lexical overlap is limited.

#### 4.1.3 Hybrid Retrieval

The hybrid retrieval system integrates sparse retrieval, dense semantic matching, and cross-encoder reranking. As shown in Table 5, this approach achieves the best performance across most metrics, demonstrating the advantages of combining lexical and semantic retrieval signals.

#### 4.1.4 Overall Retrieval Comparison

Table 6 summarizes the retrieval effectiveness of the three systems. The hybrid pipeline consistently outperforms the other approaches, indicating that multi-stage retrieval strategies can substantially improve ranking performance.

Table 5: Retrieval performance of the hybrid retrieval run.

| Metric | Score (NITA_R_JH_HY) |
|---|---|
| MAP | 0.1037 |
| MRR | 0.8955 |
| R-Precision | 0.1583 |
| nDCG | 0.2235 |
| nDCG@30 | 0.5270 |
| nDCG@100 | 0.4067 |
| P@10 | 0.6773 |
| Recall@100 | 0.1583 |

Table 6: Retrieval-only results for our submitted runs on the TREC RAG 2025 Retrieval task.

| Run | nDCG@30 | nDCG@100 | Recall@100 |
|---|---|---|---|
| bm25_NITA_JH | 0.397 | 0.300 | 0.112 |
| NITA_R_DPR | 0.486 | 0.359 | 0.133 |
| NITA_R_JH_HY | **0.486** | **0.407** | **0.158** |

### 4.2 Augmented Generation and RAG Results

Table 7 reports the performance of our Augmented Generation and full RAG submissions. The balanced precision and recall scores indicate that generated answers remain reasonably aligned with the supporting evidence segments.

### 4.3 Full RAG Task Results

The topic-level analysis presented in Table 9 indicates that system performance varies across topics depending on the availability and relevance of retrieved evidence.

### 4.4 Relevance Judgment Results

The results presented in Table 13 indicate relatively low agreement between automatically generated labels and pooled judgments, highlighting the challenges of automatic relevance assessment for complex narrative queries.

Across the retrieval experiments, the hybrid retrieval system consistently outperforms both the lexical BM25 baseline and the dense DPR retrieval model. This improvement demonstrates that combining sparse lexical matching with semantic embeddings and neural reranking can provide a stronger ranking signal than any single retrieval approach. For the generation tasks, the results indicate that the quality of retrieved evidence plays a critical role in determining the effectiveness of generated answers. Systems that retrieve more relevant and diverse evidence segments tend to achieve higher coverage and more accurate answer generation. Overall, these findings highlight the importance of robust retrieval strategies as a foundation

Table 7: Performance of our submitted runs on the TREC RAG 2025 AG/RAG task.

| Run ID | Group | Task | Weighted Precision | Weighted Recall |
|---|---|---|---|---|
| NITA_AG_JH | NITATREC | AG | 0.3888 | 0.3888 |
| ag_run_1_JH | NITATREC | RAG | 0.3888 | 0.3888 |

Table 8: Performance of our system on the TREC RAG Full Task.

| Metric | Score |
|---|---|
| Strict Vital Score | 0.19 |
| Sub Coverage | 0.36 |
| Weighted Precision | 0.472 |
| Weighted Recall | 0.472 |

Table 9: Topic-level performance for selected queries in the RAG task. SV = Strict Vital Score, SC = Sub Coverage, W.P = Weighted Precision, and W.R = Weighted Recall.

| Topic | SV | SC | W.P | W.R |
|---|---|---|---|---|
| 14 | 0.08 | 0.11 | 0.471 | 0.471 |
| 161 | 0.36 | 0.56 | 0.667 | 0.667 |
| 233 | 0.50 | 0.83 | 0.594 | 0.594 |
| 300 | 0.14 | 0.60 | 0.441 | 0.441 |
| 477 | 0.30 | 0.57 | 0.025 | 0.025 |
| 897 | 0.04 | 0.29 | 0.486 | 0.486 |

Table 10: Performance of our system on the Augmented Generation (AG) task.

| Metric | Score |
|---|---|
| Strict Vital Score | 0.18 |
| Sub Coverage | 0.35 |
| Weighted Precision | 0.481 |
| Weighted Recall | 0.481 |

for effective retrieval-augmented generation systems.

### 4.5 Discussion

Several observations regarding the behavior of retrieval and generation components within the TREC RAG 2025 framework are provided by the experimental findings. By analyzing the performance of the submitted systems in various subtasks, it is possible to gain a more comprehensive understanding of the impact of retrieval quality and evidentiary grounding on the efficacy of retrieval-augmented generation pipelines. The retrieval experiments suggest that neural retrieval methods offer distinct advantages over conventional lexical approaches. As a result of its capacity to identify direct term matches between queries and segments, the BM25 baseline functions reasonably well. Nevertheless, its efficacy is restricted when pertinent passages employ vocabulary that is distinct from the query formulation due to its dependence on surface-level lexical overlap. Conversely, the semantic similarity between passages and queries is captured by the dense retrieval model, which is based on DPR embeddings, resulting in an enhanced ranking quality. By integrating sparse lexical matching with dense semantic representations and employing cross-encoder reranking, the hybrid retrieval system further improves performance. This combination appears to more effectively balance precision and recall, resulting in an overall improvement in retrieval efficacy. The importance of evidence quality in the generation of answers is underscored by the results of the Augmented Generation (AG) task. The generation model is capable of generating coherent responses that integrate evidence-based citations when it is provided with pertinent supporting passages. Simultaneously, the observed variability across topics implies that the quality of the generated answers is significantly influenced by the comprehensiveness of the retrieved evidence. Queries with segments that address multiple aspects of the information need are more likely to attain strict vital scores and coverage metrics. In contrast, the responses that are generated frequently fail to encompass all critical aspects of the inquiry when the available evidence is limited or only partially relevant. The close relationship between retrieval and generation phases is further underscored by the results of the full Retrieval-Augmented Generation (RAG) pipeline. In this scenario, the quality of the retrieved segments is directly interdependent with the generation model. The system is capable of synthesizing multi-sentence responses that are substantiated by appropriate citations when pertinent segments are effectively retrieved. Nevertheless, the generation component is unable to mitigate for the absence of evidence when the retrieval stage fails to identify passages that are sufficiently relevant, resulting in a negative impact on the overall quality of the response. This observation underscores the significance of resilient retrieval mechanisms in end-to-end RAG systems. The challenges associated with automatically designating relevance labels for complex narrative queries are ultimately revealed by the relevance judgment results. Relevance assessment remains a challenging issue, particularly in instances where queries necessitate a nuanced interpretation of evidence passages, as evidenced

Table 11: Topic-level performance for selected queries in the AG task.

| Topic ID | Strict Vital | Sub Coverage |
|---|---|---|
| 58 | 0.46 | 0.60 |
| 161 | 0.36 | 0.56 |
| 233 | 0.50 | 0.83 |
| 213 | 0.26 | 0.25 |
| 14 | 0.04 | 0.11 |
| 200 | 0.00 | 0.00 |

Table 12: Topic-level performance for selected queries in the AG task (additional evaluation).

| Topic ID | Strict Vital | Sub Coverage |
|---|---|---|
| 58 | 0.46 | 0.60 |
| 161 | 0.36 | 0.56 |
| 233 | 0.50 | 0.83 |
| 213 | 0.26 | 0.25 |
| 14 | 0.04 | 0.11 |
| 200 | 0.00 | 0.00 |

by the comparatively low agreement between the automatically generated labels and the pooled judgments. This result implies that additional research is required to create more dependable automated relevance estimation methods for large-scale evaluation environments. The results, when considered collectively, indicate that the development of Retrieval-Augmented Generation systems is contingent upon the enhancement of retrieval efficacy. The quality and reliability of the responses that are generated are likely to be directly influenced by the refinement of reranking strategies, the improvement of evidence selection mechanisms, and the enhancement of retrieval accuracy.

## 5 Conclusion

This paper summarizes our involvement in the TREC RAG 2025 shared task, during which we created and assessed numerous systems for the Retrieval (R), Augmented Generation (AG), Retrieval-Augmented Generation (RAG), and Relevance Judgment subtasks. Our methodology investigated a variety of retrieval strategies, such as a conventional BM25 baseline, a dense retrieval model that is based on DPR embeddings, and a hybrid retrieval pipeline that integrates sparse and dense retrieval with cross-encoder reranking. Furthermore, we developed a generation pipeline that employs an instruction-tuned language model to generate evidence-based responses for the AG and RAG tasks. The significance of retrieval quality in Retrieval-Augmented Generation systems is underscored by the experimental results. The semantic representations' ability to capture relationships between queries and relevant passages was demonstrated by the dense retrieval model, which exhibited significant advances over the lexical baseline. The hybrid retrieval system demonstrated the most robust performance, indicating that the integration of semantic and lexical retrieval signals can substantially enhance the efficacy of ranking. The findings of the generation tasks indicate that language models that have been instruction-tuned can generate structured and evidence-supported responses when presented with pertinent passages. Nevertheless, the performance varied across topics, suggesting that the quality and coverage of retrieved evidence significantly affect the accuracy and completeness of the generated responses. The retrieval stage was the primary factor in the pipeline's effectiveness in the full RAG setting, underscoring the importance of high-quality retrieval in end-to-end systems. The challenges associated with automated relevance evaluation for complex narrative queries are also underscored by the relevance judgment results. However, the relatively low agreement with pooled judgments implies that automated relevance labeling is still an unresolved research issue that necessitates additional investigation. The significance of robust retrieval mechanisms in the support of effective Retrieval-Augmented Generation systems is underscored by the results of this study. In order to further improve the reliability and quality of the responses generated in large-scale information retrieval tasks, future research may investigate enhanced hybrid retrieval models, more effective reranking strategies, and advanced evidence selection techniques.

Table 13: Label agreement statistics for our submission in the Relevance Judgment subtask (qrel-event).

| Run | Group | $\kappa$ | Agree | Disagree | Tie |
|---|---|---|---|---|---|
| NITA_Qrel | NIT Agartala | 0.06 | 0.21 | 0.69 | 0.20 |

## References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Manimaran Aridoss, Khushwant Singh Bisht, and Arul Kumar Natarajan. 2024. Comprehensive analysis of falcon 7b: A state-of-the-art generative large language model. In *Generative AI: Current Trends and Applications*, pages 147–164. Springer.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *Preprint*, arXiv:2102.10073.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Support evaluation for the trec 2024 rag track: Comparing human versus llm judges. *arXiv preprint arXiv:2504.15205*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.