Leveraging Language Models for Textual Retrieval in the TREC 2025 DRAGUN Track

Andrew Sauer, Max Vanderbeek, Dr. Darren Lim

Siena University, New York

ABSTRACT

The TREC 2025 DRAGUN Track investigates question generation and evidence-based reasoning over news and media content. Our system uses the Facebook AI Similarity Search (FAISS) framework for dense document vectorization and nearest-neighbor retrieval. Each article is used to retrieve semantically related content, which is summarized using the Bidirectional and Auto-Regressive Transformer (BART) model and combined with structured ChatGPT prompts to generate candidate questions. These questions are grouped into four categories: author bias, publication reputation, factual query, and public reaction using a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) classification model. Each question is re-queried through the FAISS index to gather supporting documents, segmented with a cross encoder for fine-grained relevance filtering, and then passed to an LLM (gpt-4.1) to generate answers grounded in the source material. The resulting question–answer pairs are synthesized into coherent paragraphs, and citation alignment is performed through a secondary FAISS index. System architecture and results are presented.

1 INTRODUCTION

The TREC 2025 DRAGUN Track focuses on question generation, retrieval, and reasoning over media and news content to evaluate systems that can support media literacy and critical analysis. The track's primary objective is to assess how automated systems can generate meaningful questions about the trustworthiness of a given article and retrieve reliable, evidence-based responses from large document collections. These questions and answers aim to help readers identify potential bias, assess source credibility, and evaluate factual accuracy within online media.

This task is relevant to ongoing research in natural language understanding, information retrieval, and large language model reasoning. The ability to generate responses grounded in verifiable evidence without hallucination is increasingly critical as generative models are deployed in news and information contexts. The DRAGUN track provides a standardized framework to test and compare such systems, enabling the community to measure progress in explainable and evidence-backed text generation.

Our team's goal in participating in the 2025 Dragun Track was twofold. First, we aimed to achieve a performance level above the median score across all participating teams according to the NIST scoring. Second, we sought to design a system we felt effectively guided readers toward informed judgment about the validity and reliability of news articles. To this end, we developed a multi-stage architecture combining semantic retrieval, summarization, classification, and large language model reasoning.

2 TASK DESCRIPTION

The TREC 2025 DRAGUN Track (Detection, Retrieval, and Augmented Generation for Understanding News) is a two-task evaluation effort that supports readers in assessing the trustworthiness of online news. The two sections are (1) Question Generation, in which participants generate ten critical and investigative questions per online article about aspects such as source bias, motivation, and viewpoint diversity, and (2) Report Generation, where participants use generated questions and the corresponding retrieved evidence to produce a well-attributed report of no more than 250 words that addresses the most important questions from Task 1 to allow readers to form educated judgments.

The track uses a collection of 30 target topics (articles) sampled from the MS MARCO V2.1 (segmented) web collection (approx. 114 million segments from ~11 million documents). Evaluation is carried out by human assessors who generate their own question rubrics and supporting answers; Task 1 submissions are scored automatically for alignment to assessor questions, and Task 2 reports are manually judged based on how many assessor questions they answer with citations.

Participants may choose to submit one or both tasks; runs may be automatic or manual. Submission formats include a tab-separated file for Task 1 and a JSONL file for Task 2, with specified fields for team ID, run ID, topic ID, and responses.

3. SYSTEM DESCRIPTION

Our system was designed to combine dense semantic retrieval with summarization, question generation, classification, and evidence-grounded reasoning. The architecture follows a multi-stage pipeline intended to produce coherent, fact-based reports that support critical evaluation of news articles.

3.1 DOCUMENT STORAGE AND RETRIEVAL

To facilitate streamlined access to relevant documents, a dockerized FastAPI server was deployed on a personal Dell R720xd running Arch Linux. Onboard is 88gb of DDR3 memory and two Intel E5-2670 v2 CPUs. While outdated, the only heavily resource intensive part of the process was indexing the dataset which lends itself well to multithreading and was simply left to run overnight. A RESTful API endpoint that accepted queries submitted via Python's requests library and returned semantically relevant documents in JSON format. The API was designed with modularity in mind, enabling easy scaling and integration with downstream components. An incoming query triggers a retrieval pipeline that interfaces directly with the FAISS index, ensuring low-latency responses and consistent handling of concurrent requests.

As previously mentioned, to facilitate document retrieval the MS MARCO dataset was vectorized using FAISS. Each document was encoded into high-dimensional embeddings and stored within the index. When a target topic was submitted to the API, it was vectorized using the same embedding model and employed as a query in a k-nearest neighbors search against the index. This process yielded a set of semantically related documents, which served as the evidence base for subsequent question generation.

3.2 SUMMARIZATION AND QUESTION GENERATION

First, the retrieved documents were summarized using BART to condense information while retaining key factual details. We found this step greatly reduced the runtime

and token usage of our pipeline while having no effect on the overall quality of the final reports. These summaries were combined with structured prompts and processed through OpenAI's API using GPT-4.1 to generate a list of up to ten candidate questions for each article. Prompts were designed to encourage diversity in the question set and to cover multiple dimensions of media literacy and information credibility.

3.3 QUESTION CLASSIFICATION

Each generated question was categorized into one of four analytical groups: author bias, publication reputation, factual query, and public reaction. A fine-tuned BERT classification model was used for this task. The model was trained on a manually labeled dataset of example questions for each category, allowing the system to automatically assign the appropriate focus.

3.4 EVIDENCE RETRIEVAL AND GROUNDING

The FAISS index was then queried again, this time with the classified questions, to retrieve the top-n relevant documents. Through experimentation we found that n=3 offered a good balance of speed and information. These documents were segmented by sentence, and a cross-encoder model was used to score and select the most relevant segments from each topic for each question. The filtered segments were supplied to the LLM together with the original question to generate a response grounded in the retrieved evidence. The prompt specified that an output should only be produced if sufficient context was available from the retrieval segments and to indicate when a question could not be answered from the provided information.

3.5 REPORT GENERATION AND CITATION ALIGNMENT

The resulting question–answer pairs were combined to form a concise, coherent report for each target article. The report text was then segmented at the sentence level and re-indexed into a small FAISS instance for citation alignment. Each sentence was matched against its most relevant supporting source to ensure that all statements were traceable to verifiable evidence within the retrieved corpus.

The code was implemented with a focus on modularity and interpretability. Each stage could be independently evaluated and refined, enabling transparent inspection of both intermediate and final outputs.

4. RESULTS

Metric	Mean Score	Our Score
Supportive	0.108398158	0.125238427
Contradictory	0.002670940	0.0053968

Table 1

The key evaluation metrics for our generated reports are summarized in Table 1, which lists both the scores assigned to our system by NIST and the median scores across all submitted runs.

The generated reports were evaluated by NIST using the official DRAGUN supportive and contradictory scoring framework. These metrics quantify how well a report covers the answer nuggets defined by the NIST assessors and how often it produces content that conflicts with those nuggets. Higher supportive scores indicate stronger factual coverage, while lower

contradictory scores indicate fewer factual inconsistencies.

Two sets of numbers were used in the analysis: (1) the scores NIST assigned to our system for each topic, and (2) the median scores across all submitted runs. The medians serve as a field-wide baseline and provide context for interpreting our system's output quality. The key results are summarized in Table 1.

4.1 SUPPORTIVE SCORES

Across all evaluated topics, the supportive score assigned to our system averaged approximately 0.12524 (see Table 1). This value reflects the proportion of answer nuggets that were either fully supported or partially supported in the report, weighted by the importance of each underlying question. Scores in this range indicate that the system consistently captured a meaningful portion of the expected evidence, with a mix of full and partial coverage across topics.

The median supportive score across all participating systems was approximately 0.10839 (see Table 1). Topic-level medians showed substantial variability, ranging from near zero to more than 0.22. Higher medians correspond to articles where the assessors' expectations were clearer or easier for systems to recover from the available evidence. Lower medians were associated with articles that were harder to fact-check or required synthesis across more diffuse sources. Our system performed above the global median, indicating that it retrieved and presented relevant factual material more reliably than the typical participating run.

4.2 CONTRADICTORY SCORES

NIST assigned contradictory scores to our system that averaged 0.00539 across topics (see Table 1), with most topics at or near zero. This score reflects the proportion of answer nuggets for which the generated report contradicted the assessor-defined facts. A near-zero score indicates that the system did not introduce assertions that conflicted with the ground-truth nuggets.

The median contradictory score across all submitted runs was approximately 0.00267 (see Table 1), confirming that most systems—including ours—produced reports that avoided direct factual contradictions. Our results fall cleanly within this pattern, showing that the grounding and alignment stages effectively constrained the model from generating factually incorrect claims.

4.3 INTERPRETATION

Taken together, the results show a system that produces factually conservative outputs: it avoids contradictions and supplies partial or full support for a sizable share of the rubric-defined evidence. The supportive scores (see Table 1) indicate meaningful—but not complete—coverage of the expected facts across topics. The near-zero contradictory scores (see Table 1) confirm that the retrieved evidence and grounding mechanisms prevented the model from fabricating or misrepresenting key points.

Relative to the median performance across all submitted runs, the system achieved above-average supportive coverage while matching the field-wide strength in minimizing contradictions. These findings highlight that the retrieval, grounding, and citation-alignment pipeline produced factually aligned summaries while keeping the risk of misinformation low. Subsequent refinements will focus on expanding supportive coverage without sacrificing this stability.

5. CONCLUSION

In this work, we presented a multi-stage system for evidence based report generation within the TREC 2025 DRAGUN Track. The system integrates dense semantic retrieval, abstractive summarization, question generation, classification, and grounding using large language models, with an emphasis on citation alignment and traceability of all statements.

Evaluation by NIST demonstrates that the system reliably produces factually conservative reports. Supportive scores indicate that a meaningful portion of the rubric-defined evidence is captured, while contradictory scores remain near zero, confirming that the model avoids introducing claims that conflict with verified sources. Comparison to median scores across all submitted runs shows that the system exceeds the typical level of supportive coverage while matching the field-wide strength in minimizing contradictions.

These results highlight the effectiveness of combining retrieval, grounding, and structured prompting to produce reliable, evidence-based summaries. The modular pipeline allows for transparent inspection and independent refinement at each stage, supporting future improvements in both coverage and interpretability.

Future work will focus on increasing the proportion of fully supported answer nuggets, improving cross-document reasoning, and further refining the alignment between generated content and source evidence. The findings presented here underscore the potential of integrating dense retrieval and large language model reasoning to produce trustworthy, actionable insights from complex media content.

6. REFERENCES

1. *Bart Large CNN* (no date) *facebook/bart-large-cnn · Hugging Face*. Available at: https://huggingface.co/facebook/bart-large-cnn (Accessed: 23 October 2025).

2. Devlin, J. *et al.* (no date) *Bert*, *BERT - transformers 3.0.2 documentation*. Available at: https://huggingface.co/transformers/v3.0.2/model_doc/bert.html (Accessed: 01 December 2025).

3. *FAISS documentation* (no date) *Welcome to Faiss Documentation - Faiss documentation*. Available at: https://faiss.ai/index.html (Accessed: 03 November 2025).

4. Zhang, D., Smucker, M.D. and Clarke, C.L.A. (no date) *Overview of the TREC 2024 lateral reading track*. Available at: https://trec.nist.gov/pubs/trec33/papers/Overview_lateral.pdf (Accessed: 22 November 2025).