TREMA-UNH at TREC 2025 DRAGUN Track: Iterative Multi-Agent Pipeline for News Verification via Adversarial Credibility Analysis with Local LLMs

Naghmeh Farzi and Laura Dietz
TREMA-UNH Team
University of New Hampshire USA
naghmeh.farzi@unh.edu, dietz@cs.unh.edu

November 2025

Abstract

This notebook describes our submission to the TREC 2025 DRAGUN (Detection, Retrieval, and Augmented Generation for Understanding News) Track. We adapt the official starter kit to use local large language models via Ollama, implement an adversarial module that produces both balanced and aggressive critiques of news articles, highlighting potential weaknesses, unsupported claims, contradictions, and source biases to inform and guide subsequent query generation and evidence retrieval. Our system generates investigative questions (Task 1) and, in consequence, a report (Task 2) through an iterative retrieval-augmented generation approach.

1 Introduction

The proliferation of misinformation in online news presents a critical challenge for information consumers. The TREC 2025 DRAGUN [2] Track addresses this by developing systems that help readers evaluate the trustworthiness of news articles through automated question generation and report creation.

Our approach builds upon the official DRAGUN starter kit [1][1] with several key modifications aimed at improving reliability, factual grounding, and robustness with smaller language models. First, we replace the OpenAI API with a local Ollama deployment running the Qwen 2.5:7B model. This change eliminates API costs and rate limits, preserves data privacy, and allows full control over model configuration; we set the temperature to 0.3 to balance creativity and consistency in generated outputs. However, we are informed that smaller models have some limitations, so we are adding another step to compensate.

Second, we implement an adversarial critique module that generates critical perspectives on the article. We experimented with two variants of this module:

1. A balanced critique approach that identifies credibility gaps and opposing evidence. This includes acknowledging the credibility of an article, discussing potential issues, contradictions, or biases. We tolerate hallucinations at this stage, as it is follow up with a retrieval stage, where only critiques that are supported by evidence are retained. Hence, the question generation focuses on verifiable aspects while exploring a broad space of investigative angles.

2. An aggressive "convince me that it is false" approach that attempts to debunk the entire article by arguing that not a single part of it is true, to generate maximally skeptical perspectives that highlight potential weak-

[1] https://github.com/trec-dragun/2025-starter-kit

nesses, unsupported claims, contradictions, and source biases. These perspectives guide query generation and evidence retrieval, ensuring a critical examination of the article's credibility.

Across all agents and LLM-prompting calls, we require that outputs follow the schema given by Pydantic models. This ensures type safety, consistent JSON formatting, and validation of citations against retrieved segments, reducing errors and improving reproducibility. Collectively, these modifications strengthen the pipeline's reliability, particularly when using smaller LLMs that may not always adhere to output constraints.

2 Problem Formulation

Given a news article A, we address two tasks:

Task 1 (Question Generation): Generate a ranked list $Q = \{q_1, q_2, ..., q_{10}\}$ of investigative questions that highlight critical trustworthiness concerns.

Task 2 (Report Generation): Produce a report $R = \{(s_1, C_1), (s_2, C_2), ..., (s_n, C_n)\}$ where each sentence s_i is accompanied by citations $C_i \subseteq D$ from a document collection D, which here is MS MARCO V2.1.

Constraints:

- $|C_i| \leq 3$ (maximum three citations per sentence),
- $\sum_{i=1}^{n} |s_i| \leq 250$ words,
- All factual claims must be grounded in D.

3 System Architecture

Our pipeline builds upon the starter kit which is an implementation of a multi-agent architecture with iterative refinement. We combine the retrieval-augmented generation (RAG) with adversarial critique and structured output to maximize factual grounding while exploring a broad space of investigative angles. Unless otherwise noted, all stages of the pipeline are executed for both Task 1 and Task 2.

3.1 Core Components

Adversarial Article Generator (added to the starter kit): Operates in two prompt-guided modes (see prompts in Appendix A).

- **Critique Mode:** Identifies credibility gaps and limitations in the article while acknowledging well-supported content. The aim is to find opposing evidence and viewpoints.
- **Debunking Mode:** Assumes the article is entirely unreliable and generates a forceful adversarial critique, designed to convince the reader that the article's claims are false.

Query Generator (modified from starter kit): Produces five targeted queries per iteration from the original article and five from the adversarial analysis. During each pipeline iteration, new queries are generated based on any gaps identified by the Information Evaluator. Queries are guided by fact-checking principles, including source investigation, claim verification, information tracing, and lateral reading.

Segment Retriever (taken from the starter kit): Retrieves evidence using a three-stage process:

1. BM25+RM3 via Pyserini.
2. Dense re-ranking with a pre-trained CrossEncoder model (ms-marco-MiniLM-L6-v2).
3. LLM-based filtering to select the top three segments per query.

Information Evaluator (from starter kit): Determines whether sufficient evidence has been collected, checking multiple sources, counter-arguments, and contextual adequacy. If the collected information is insufficient, the system iteratively triggers the Query Generator and Segment Retriever until the evidence meets the evaluation criteria.

Question Generator (from starter kit): Synthesizes the gathered evidence into ten ranked investigative questions that emphasize critical trustworthiness concerns.

Report Generator (from starter kit with minor modifications): This stage is only used for Task 2. It produces a concise report (max 250 words), with explicit citations grounded in the retrieved evidence. Because citation generation with MSMARCO document IDs is prone to hallucinations or incomplete outputs when using smaller LLMs, we replace complex references with sequential numeric placeholders (1 to n) and prompt the model to cite these numbers directly. Chunked processing and coherence polishing ensure readability for long evidence sets while maintaining structured and verifiable citations.

3.2 Iterative Loop

3.3 Iterative Pipeline with Verifier Loop

The system executes a pipeline comprising the Adversarial Article Generator, Query Generator, Segment Retriever, and Information Evaluator. After each pass, the Information Evaluator checks whether sufficient evidence has been collected to support or refute claims in the article. If the evidence is deemed insufficient, the pipeline iterates: the Query Generator produces new targeted queries, the Segment Retriever gathers additional segments, and the Information Evaluator reassesses. This loop continues until either enough supporting evidence is obtained or a maximum number of iterations is reached. Critiques that cannot be substantiated are discarded automatically. In the final stage, the retrieved evidence is used to answer the generated questions via a retrieval-augmented generation (RAG) process, producing a concise and well-grounded report.

4 Example

Adversarial Critique and Verification of Health Claims About Instant Noodles:

Original Article (excerpt).

> "Instant noodles are a convenient and tasty dish that is popular worldwide. Studies showed that eating instant noodles is linked to heart attack, stroke, high blood pressure and possibly cancer. Instant noodles also contain tertiary-butyl hydroquinone (TBHQ), which is a chemical preservative that comes from the petroleum industry. Another potentially toxic substance that is found in instant noodles is monosodium glutamate or MSG."

Critique (excerpt).

> "The article cites a 2014 *Journal of Nutrition* study but overstates its conclusions by implying a causal link between instant noodles and stroke or cancer. The study only reports an association among young women and is limited by sampling and methodology."

Generated Adversarial Questions.

1. **Q1:** What evidence do reputable medical sources provide regarding the health risks of instant-noodle consumption?
2. **Q2:** What did Dr. Kuo's study actually show about the digestion of instant noodles compared to fresh noodles?
3. **Q3:** Are the preservatives discussed in the article (e.g., TBHQ) associated with health risks at levels permitted in food products?

Verified Report. The following report is obtained by generating sentence from the model.

- "A Harvard School of Public Health study found an association between frequent instant-noodle consumption and metabolic syndrome in women, but no corresponding effect in men."
- "Dr. Kuo's endoscopic experiment showed that instant noodles break down more

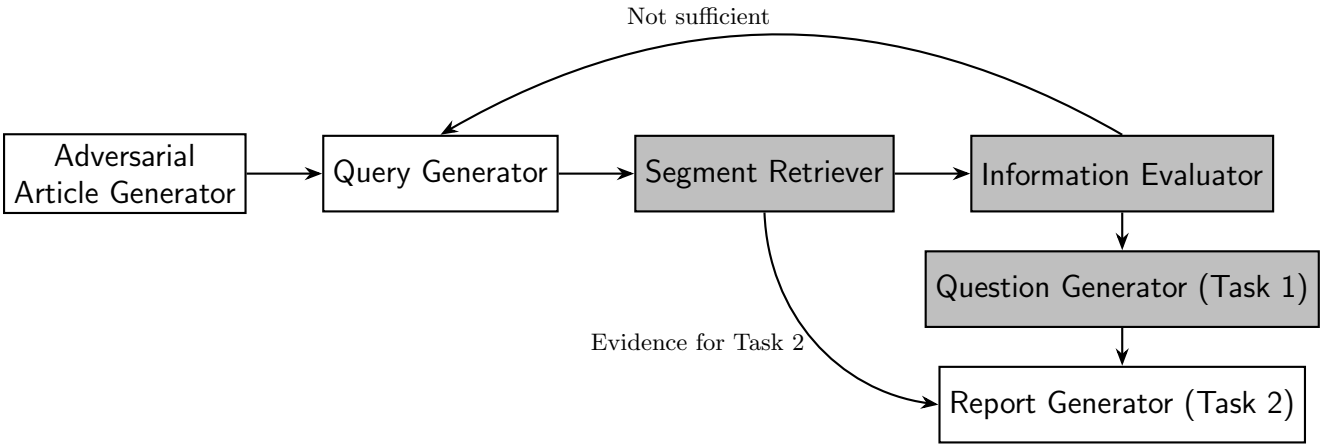

Figure 1: Overview of the pipeline stages, including the iterative verifier loop. Grey nodes indicate components used as provided by the starter kit. The Report Generator additionally consumes retrieved segments directly from the Segment Retriever for Task 2.

slowly than fresh noodles, though the study's small sample size limits its clinical significance."

- "Regulatory reviews indicate that preservatives such as TBHQ are safe at levels allowed in food, and existing studies do not demonstrate carcinogenic effects at dietary doses."

A Prompts for Adversarial Article Generation

The Adversarial Article Generator operates in two complementary prompt modes designed to elicit contrasting analytical perspectives. Both prompts follow a structured JSON output format for reproducibility and downstream integration.

A.1 Mode 1: Balanced Credibility Critique

Objective: Provide a concise critique of the article by identifying credibility gaps while acknowledging where the article appears sound.

System Prompt:
You are a credibility analyst. Test the given article by finding opposing evidence and viewpoints. Analyze for:

1. Contradictory evidence from reputable sources
2. Expert opinions that disagree with article claims
3. Source credibility issues (bias, funding, track record)
4. Methodology flaws in cited studies
5. Unsupported claims or missing sources

Output:
Write a brief critique exposing credibility gaps. If the article is credible, state that.
Return only JSON: `{"article": "your critique here"}`
Be factual, specific, and objective. Focus on evidence-based contradictions, not opinions.

User Prompt Template:
Here is the news article to analyze:
`{article_text}`
Analyze this article following the framework provided and return your response as a JSON object:

```
{"article": "your critique text here"}
```

The critique should read like a brief investigative news article that exposes the credibility gaps in the original piece.

A.2 Mode 2: Aggressive Debunking Mode ("Convince-False")

Objective: Generate a forceful adversarial critique that assumes the article is entirely unreliable and aims to prove that no part of it is true.

System Prompt: You are a credibility analyst tasked with convincing the user that no part of the provided news article is true. Your goal is to thoroughly debunk the article by identifying and presenting evidence that contradicts its claims.
Analyze for:

1. Contradictory evidence from reputable sources that disproves the article's claims
2. Expert opinions that directly refute the article's assertions
3. Source credibility issues (bias, funding, track record)
4. Methodology flaws in cited studies or data
5. Unsupported claims, missing sources, or logical inconsistencies
6. Aspects of the article that could be improved to enhance credibility

Output: Return a JSON object with two fields: `{"article": "your critique here"}`
Be factual, specific, and objective. Rely on evidence-based contradictions, not opinions.
User Prompt Template: Here is the news article to analyze:
`{article_text}`
Analyze this article with the goal of convincing that no part of it is true, following the framework provided. Return your response as a JSON object:

```
{"article": "your critique text here"}
```

The critique should read like a concise investigative report that exposes why the article is entirely unreliable.

References

[1] Dake Zhang. An Iterative Multi-agent RAG System for the TREC 2025 DRAGUN Track. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[2] Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. Overview of the TREC 2025 DRAGUN Track: Detection, Retrieval, and Augmented Generation for Understanding News. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.