A LangChain-Based Framework for Investigative Question Generation Using Large Language Models

Adnan Faisal
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
Chattogram, Bangladesh
ajfaisal1208023@gmail.com

Shiti Chowdhury
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
Chattogram, Bangladesh
shitichowdhury21@gmail.com

Abstract—The increasing prevalence of online misinformation has amplified the demand for automated approaches that assist readers in assessing the credibility of news articles. The TREC 2025 DRAGUN (Detection, Retrieval and Augmented Generation for Understanding News) Track addresses this need through its Question Generation task, which requires systems to formulate ranked investigative questions that support reader-oriented credibility assessment. This study presents a LangChain-based pipeline for generating focused and investigative questions from news articles in the MS MARCO V2.1 segmented corpus. The proposed framework combines structured prompt design, controlled decoding and semantic reranking to improve question relevance, coherence and interpretability. We have evaluated several experimental configurations covering Qwen-based, Mistral-based and reasoning-oriented large language models using the rectified DRAGUN evaluation protocol, where compound questions are removed prior to scoring. Our experimental results indicate that reasoning-aligned models exhibit stronger and more consistent performance under strict evaluation constraints, with the CUET-QwQ-32B configuration achieving the highest average score among our submissions. At the same time, Qwen-14B variants demonstrate stable and competitive performance across diverse topics, showing substantial agreement with assessor-defined evaluation rubrics. Overall, our findings demonstrate that a structured and modular question-generation pipeline can effectively translate large language model reasoning into practical support for reader-centric news trustworthiness assessment, while also providing insights for extending such systems toward multi-source report generation.

Keywords

Question Generation, News Credibility Assessment, LangChain Pipeline, Semantic Reranking, Large Language Models (LLMs), TREC DRAGUN

I. Introduction

The widespread dissemination of digital media has significantly complicated the assessment of online news credibility. Readers are routinely exposed to misleading claims, selective framing and biased narratives that can shape public opinion in subtle ways. Conventional fact-checking approaches generally operate under the assumption that specific claims and supporting evidence are already identifiable. In open-web settings, however, the primary challenge lies in identifying which questions must be asked before meaningful verification can take place.

The TREC 2025 DRAGUN (Detection, Retrieval and Augmented Generation for Understanding News) Track, organized by the National Institute of Standards and Technology (NIST), addresses this challenge through its Question Generation task (Task 1) [2]. This task requires systems to generate ranked, investigative questions that assist readers in evaluating the trustworthiness of a news article. Unlike verdict-oriented fact-checking systems, DRAGUN adopts a reader-centric approach that encourages deeper examination of source credibility, contextual gaps and potential bias through carefully formulated investigative questions.

Recent advances in large language models (LLMs) have led to substantial progress in automated question generation. Varifocal QG [1] has employed multi-focal reasoning to produce context-aware questions from diverse semantic perspectives, while AutoQGS [4] has introduced adaptive auto-prompting techniques to increase question diversity in low-resource settings. Despite these improvements, many existing approaches rely primarily on prompt diversity and lack mechanisms for iterative reasoning and structural control. The Progressive Thought Refinement (PTR) framework [7] has addressed this limitation by enabling LLMs to iteratively refine their reasoning through weighted thought masking, making it well suited for investigative question generation tasks.

Building on these developments, we propose a LangChain-based multi-stage pipeline that integrates structured prompting, controlled decoding and semantic reranking to generate ranked investigative questions for news credibility assessment task. Our system is evaluated using the rectified DRAGUN evaluation protocol, in which compound questions are automatically detected and removed prior to scoring [2]. Results across multiple model configurations indicate that reasoning-aligned and structurally constrained generation strategies yield more consistent performance under strict evaluation criteria than approaches that rely solely on model scale. In particular, reasoning-oriented models demonstrate greater robustness to compound-question filtering, while Qwen-based variants provide stable and competitive performance across a wide range of topics. These observations high-

light the importance of atomic question formulation and structured generation pipelines for effective reader-centric trustworthiness assessment.

- We have implemented a LangChain-based, multi-stage question generation framework that systematically generates ranked investigative questions for evaluating news trustworthiness through the integration of prompt structuring, decoding control and relevance-based reranking.
- We have incorporated Progressive Thought Refinement (PTR) principles to enable iterative reasoning, resulting in improved question focus, structural consistency and contextual alignment under strict evaluation constraints.

For implementation details and access to the full codebase, refer to the GitHub repository: https://github.com/AJFaisal002/Investigative-Question-Generation-IQG.

II. Related Works

Question Generation (QG) is a foundational component of fact-checking and news credibility analysis, as it helps uncover missing information and guides subsequent evidence retrieval. Early neural QG models [5], [6] were largely developed for reading comprehension and relied on answer-aware formulations over predefined passages, which restricted their effectiveness in open-domain environments.

The emergence of large language models (LLMs) has significantly broadened the capabilities of automated question generation. Varifocal QG [1] has proposed multi-focal reasoning to capture complementary semantic perspectives, while AutoQGS [4] has introduced adaptive auto-prompting to enhance question diversity under limited supervision. Progressive Thought Refinement (PTR) [7] has further advanced this line of work by enabling iterative reasoning for improved contextual accuracy, a trend has also emphasized in recent surveys on reasoning-centric Question Generation (QG) [8].

Beyond standalone generation, retrieval-augmented approaches have explored tighter integration between retrieval and question formulation. Keller et al. [9] has employed diversified segment selection with MMR-based reranking to improve relevance, while Fu et al. [10] has investigated hybrid Retrieve–then–Generate pipelines for conversational settings.

Within this evolving research landscape, the TREC 2025 DRAGUN Task 1 establishes investigative question generation as a reader-centric mechanism for assessing news trustworthiness, supported by a standardized evaluation protocol that emphasizes question quality and structural compliance [2]. Motivated by this task, our work adopts a LangChain-based multi-stage framework that integrates structured prompt engineering, constrained generation and semantic relevance assessment to produce ranked investigative questions, building upon baseline systems introduced for the DRAGUN track [3].

TABLE I: Topic Schema in trec-2025-dragun-topics.jsonl

Field	Description
docid	Document identifier corresponding to an entry in the MS MARCO V2.1 segmented collection.
url	Web address of the original source article.
title	Article headline capturing the central theme or claim.
headings	Extracted section titles used to preserve the article's structural layout.
body	Full article text segmented and preprocessed for downstream question generation.

III. Tasks

A. Task Description

The dataset used in the TREC 2025 DRAGUN Track is released through the official starter kit in the file trec-2025-dragun-topics.jsonl. It contains 30 target news articles (topics), each encoded as a structured JSON object with predefined fields describing the article content and associated metadata, as outlined in Table I. All topics are sampled from the MS MARCO V2.1 (Segmented) collection, which comprises over 114 million text segments derived from approximately 11 million web documents.

The topics span diverse domains, including politics, science, health and culture. Each topic provides sufficient context for investigative question generation while preserving uncertainty around key facts and sources.

As an illustrative example, the topic "The Greek-Canadian Origins of the Hawaiian Pizza" focuses on the historical development of Hawaiian pizza, naturally encouraging questions related to authenticity, cultural attribution and factual consistency.

IV. Methodology

A. System Overview

Our approach to Task 1: Question Generation in the TREC 2025 DRAGUN Track follows a three-stage pipeline consisting of a Retriever, a Generator and a Reranker. The system is implemented within the LangChain framework, enabling modular design, ease of experimentation and consistent execution across multiple runs.

The Retriever is responsible for preprocessing and organizing the input news articles, while the Generator employs large language models (LLMs) to produce candidate investigative questions. Subsequently, the Reranker evaluates and orders these questions by emphasizing semantic diversity and contextual relevance. This modular design facilitates iterative refinement and supports efficient generation of ranked questions for multi-topic news trustworthiness assessment.

B. Retriever Module

The Retriever module constructs structured inputs for the question generation stage by extracting the title and

body fields from each article in the trec-2025-dragun-topics.jsonl dataset.

The extracted text undergoes light preprocessing, including the removal of HTML markup, non-informative symbols and excessive whitespace, before being combined into a unified prompt representation.

No external retrieval mechanisms are incorporated, ensuring that generated questions remain fully grounded in the source article and mitigating the risk of hallucination or contextual drift.

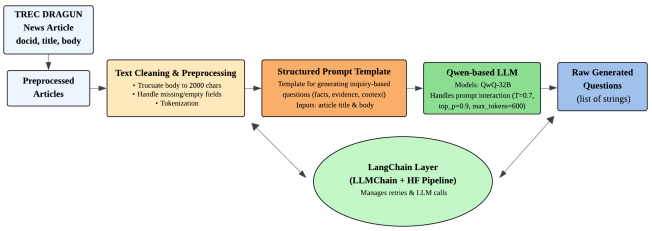

Fig. 1: Retriever pipeline for preprocessing and prompt preparation using TREC DRAGUN articles.

As illustrated in Fig. 1, the Retriever module preprocesses news articles from the TREC DRAGUN dataset by applying basic text normalization, filtering empty or missing fields and truncating article bodies to a maximum of 2,000 characters. The processed title and body contents are then assembled into a structured prompt and routed through the LangChain framework to the QwQ-32B model for investigative question generation.

C. Generator Module

The Generator constitutes the core reasoning module of the system and is realized using LangChain's LLMChain abstraction. Given a structured prompt produced by the Retriever, the selected large language model is instructed to generate ten ranked investigative questions that emphasize credibility assessment, potential bias and diverse viewpoints expressed in the article.

A concise prompt template is used to guide the generation process:

Prompt Template

"You are a professional fact-checker and investigative media literacy expert, trained on resources like PolitiFact, MBFC, NewsQA and FEVER.

Your task: Generate 10 sharp, critical and investigative questions a thoughtful reader should ask to assess the trustworthiness of a given news article. These must be specific to the article, ranked from most to least important for evaluating credibility."

Decoding parameters (temperature = 0.6, top-p = 0.9, max_tokens = 600) are chosen to balance generative diversity and coherence while maintaining factual consistency. This configuration supports the production of focused and analytically grounded questions under the rectified evaluation protocol.

As shown in Fig. 2, the Generator module employs the QwQ-32B model within the LangChain pipeline to perform investigative question generation. The ranked outputs produced at this stage are subsequently forwarded to the Reranker for refinement based on relevance, diversity and structural compliance.

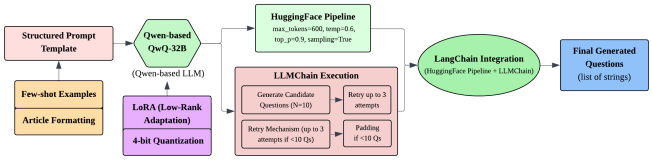

Fig. 2: Generator pipeline for investigative question generation using QwQ-32B and LangChain integration.

D. Reranker Module

The Reranker module refines the generated question set by promoting semantic diversity and ensuring contextual relevance. Using the Qwen3-Reranker-8B model, pairwise cosine similarity between question embeddings is calculated as:

$$S(q_i, q_j) = \frac{E(q_i) \cdot E(q_j)}{\|E(q_i)\| \, \|E(q_j)\|} \tag{1}$$

Questions for which $S(q_i, q_j) \geq 0.8$ are discarded to reduce redundancy and overlap in semantic content.

To prioritize analytically informative questions, we apply a weighted relevance scoring mechanism that emphasizes bias detection, credibility assessment and source reliability.

$$R(q_i) = \sum_{k=1}^{N} w_k \times \mathrm{Sim}(q_i, r_k) \tag{2}$$

This reranking procedure yields a final set of ten diverse, well-formed questions per topic that align closely with the official evaluation criteria.

As shown in Fig. 3, the Reranker module applies a series of post-processing operations to the generated question list. These operations include duplicate removal, filtering of overly short or off-topic questions, completion of partial outputs and standardized formatting of the final ranked questions for evaluation.

E. Compound Question Handling

The DRAGUN Question Generation task requires each system output to be a single, atomic investigative question [2]. In practice, we have observed that large language models may occasionally produce compound questions that combine multiple inquiries (e.g., "What is X and what is Y?"), violating this constraint.

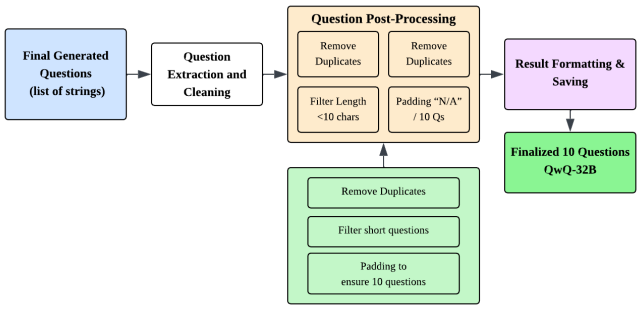

Fig. 3: Reranker pipeline for question filtering, post-processing and final selection.

While our reranking module follows the DRAGUN starter-kit baseline [3] and focuses on relevance and diversity, it does not explicitly enforce atomicity. The updated DRAGUN evaluation therefore introduces an automated filtering step using gpt-oss-120b to remove compound questions prior to scoring [2]. This rectified evaluation underscores the importance of enforcing atomic question structure in future systems.

V. Further Settings and Implementation Details

All experiments have been conducted on Kaggle using an NVIDIA P100 GPU with Python 3.10. The overall system has been implemented using the LangChain framework, enabling modular design, reproducibility and structured orchestration across all components.

Final Experimental Configuration

Framework:
LangChain with FP16 precision and a 4,096-token context window
Models: QwQ-32B (Generator), Qwen3-Reranker-8B (Reranker)
Generation Parameters: max_new_tokens = 600, temperature = 0.6, top_p = 0.9
Reranking Criterion: Cosine similarity threshold of 0.8
Outcome: Balanced and robust performance under the rectified DRAGUN evaluation protocol.

The final configuration employs QwQ-32B as the primary generator for investigative question production, while Qwen3-Reranker-8B is used for semantic filtering and ranking. Generation employs controlled decoding to balance diversity and coherence, followed by cosine similarity–based reranking to mitigate redundancy and ensure robustness under automated evaluation.

VI. Results and Comparative Analysis

Table II represents the performance of all ten submitted runs evaluated under the official NIST rubric for the TREC 2025 DRAGUN Task 1. Scores are calculated as weighted similarity scores between the system-generated questions and the assessor-defined rubric questions.

A. Result Analysis

The experimental results show consistent performance trends across the submitted runs under the rectified evaluation. CUET-QwQ-32B achieved the highest average score (0.215), demonstrating strong robustness to compound-question filtering. Among Qwen-based models, CUET-qwen14B-v2 performed best, balancing investigative depth and structural consistency. Higher temperature values has increased diversity but reduced factual grounding, while the smaller Qwen3-4B model has produced fluent yet shallow questions. Overall, moderate decoding settings (temperature=0.5, top-p=0.8) has resulted in the most stable performance, reinforcing the effectiveness of the proposed LangChain-based pipeline for DRAGUN Task 1.

B. Decoding Sensitivity and Comparative Evaluation

Figure 4 shows the performance trend across all CUET runs under the rectified DRAGUN Task 1 evaluation. Higher-capacity and reasoning-oriented models consistently outperform smaller configurations, indicating the importance of model capacity and stable decoding strategies. Moderate decoding settings have yielded stable outputs, whereas aggressive sampling has degraded rubric alignment.

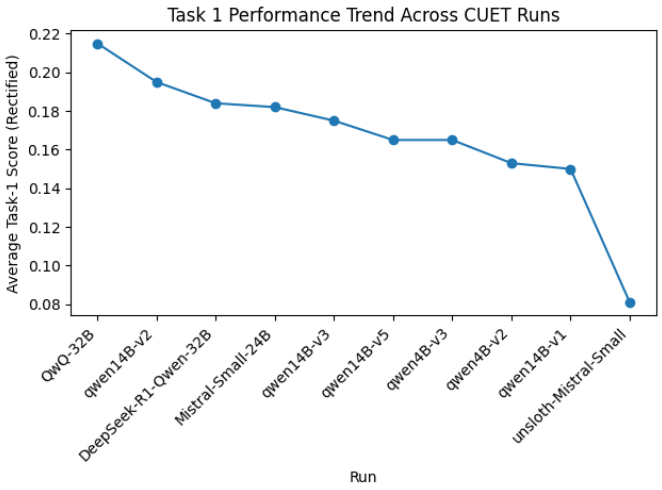

Fig. 4: Task 1 performance trend across CUET runs under the rectified evaluation.

Figure 5 compares the submitted runs based on their average Task 1 scores. CUET-QwQ-32B has achieved the highest performance, demonstrating strong robustness to compound-question filtering. Among the Qwen-based systems, configurations incorporating decoding optimization and LoRA-based fine-tuning with quantization have achieved competitive results. In contrast, smaller models have produced fluent but comparatively shallow questions.

TABLE II: CUET Task 1 (Question Generation) runs and results under the rectified TREC 2025 DRAGUN evaluation.

Run ID	Model	Temp	Top-p	Avg. Score	Prompt Style	Remarks
CUET-QwQ-32B	QwQ-32B	0.6	0.9	0.215	Ranked	Strongest
CUET-qwen14B-v2	Qwen3-14B	0.5	0.8	0.195	Balanced	Best Qwen-14B
CUET-DeepSeek-R1-Qwen-32B	DS-R1 + Qwen-32B	0.6	0.9	0.184	Reasoning	Robust
CUET-Mistral-Small-24B	Mistral-24B	0.6	0.9	0.182	Compact	Topic-sensitive
CUET-qwen14B-v3	Qwen3-14B	0.7	0.9	0.175	Broad	High diversity
CUET-qwen4B-v3	Qwen3-4B	0.5	0.8	0.165	Structured	Limited reasoning
CUET-qwen14B-v5	Qwen3-14B	0.6	0.9	0.165	Reranked	Low redundancy
CUET-qwen4B-v2	Qwen3-4B	0.7	0.9	0.153	Exploratory	Low alignment
CUET-qwen14B-v1	Qwen3-14B	0.3	0.85	0.150	Concise	Shallow output
CUET-unsloth-Mistral-Small	Unsloth-Mistral	0.6	0.9	0.081	Lightweight	Penalized

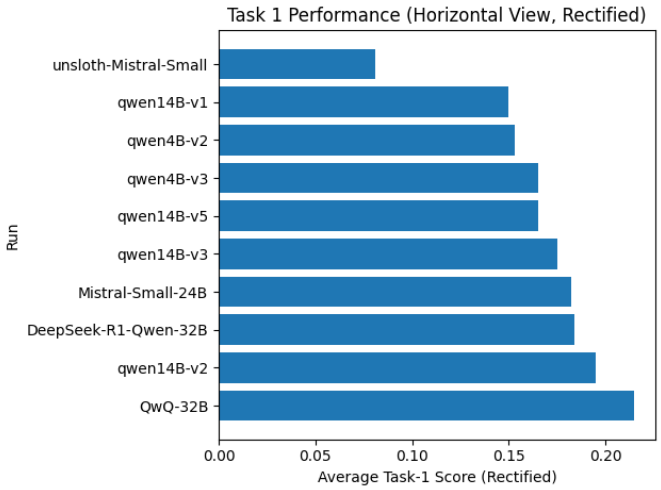

Fig. 5: Rectified Task 1 performance comparison across CUET model variants.

Overall, decoding calibration and parameter-efficient fine-tuning (e.g., LoRA) are key to stable investigative question generation in DRAGUN Task 1.

VII. Ablation Study

An ablation study has been performed to analyze the impact of key components in the proposed LangChain-based pipeline. Removing Progressive Refinement (PR) has reduced contextual diversity, while excluding the Semantic Reranker has increased redundancy and weakened rubric alignment. Disabling Controlled Decoding using aggressive sampling values (temperature=0.9, top-p=0.95) has improved linguistic variation but degraded factual consistency.

Overall, semantic reranking and controlled decoding are key to stable, rubric-aligned question generation, with the full pipeline outperforming all ablated variants. Table III summarizes the observed trends.

TABLE III: Ablation Analysis under the Rectified DRAGUN Evaluation

Configuration	Relative Score Change	Redundancy
Full Pipeline	Baseline	Low
Without Semantic Reranker	↓ Moderate	High
Without Progressive Refinement	↓ Mild	Medium
Without Controlled Decoding	↓ Moderate	Medium

VIII. Conclusion and Future Work

This study has presented a LangChain-driven multi-stage framework for investigative question generation in the TREC 2025 DRAGUN Task 1. By combining structured prompting, guided generation and semantic reranking, the system has consistently generated context-aware and rubric-aligned questions across diverse news topics. Results highlight the role of controlled decoding and progressive refinement in achieving a stable balance between creativity, coherence and credibility.

Future work will focus on integrating retrieval-augmented grounding to further improve factual accuracy and dynamic self-evaluation mechanisms to further enhance adaptive reasoning in complex news scenarios. We also aim to explore lightweight reasoning models and efficiency-oriented optimizations to support scalable deployment in real-world news analysis settings.

IX. Ethical Statement

This study upholds established ethical principles in artificial intelligence research and responsible data handling. All experiments were performed exclusively on the publicly released dataset of the TREC 2025 DRAGUN Track, ensuring full adherence to its data usage guidelines and open research standards. No personally identifiable or sensitive information was collected or processed during the study. The generated outputs from large language models (LLMs) were carefully examined to mitigate bias, maintain factual integrity and prevent the dissemination of misleading or harmful content.

X. Acknowledgement

The authors thank the TREC 2025 DRAGUN Track organizers and the National Institute of Standards and Technology (NIST) for providing the framework and evaluation resources and acknowledge the open-source contributions of LangChain and HuggingFace that supported this work.

References

[1] N. Ousidhoum, S. Pannitto, Y. G. Choi and M. Ballesteros, "Varifocal Question Generation for Fact-Checking," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 3241–3256.

[2] D. Zhang, M. D. Smucker and C. 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), National Institute of Standards and Technology (NIST), 2025.

[3] D. Zhang, "An Iterative Multi-Agent RAG System for the TREC 2025 DRAGUN Track," in The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025), National Institute of Standards and Technology (NIST), 2025.

[4] W. Xiong, S. Zhao and J. Liu, "AutoQGS: Auto-Prompt for Low-Resource Knowledge-Based Question Generation," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 11842–11855.

[5] X. Du, J. Shao and C. Cardie, "Learning to Ask: Neural Question Generation for Reading Comprehension," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017, pp. 1342–1352.

[6] Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao and M. Zhou, "Neural Question Generation from Text: A Preliminary Study," in Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP), 2017, pp. 662–671.

[7] X. Du, C. Liu and T. Zhang, "Think Thrice Before You Act: Progressive Thought Refinement for Large Language Models," arXiv preprint arXiv:2402.07119, 2024.

[8] Y. Guo, H. Wang and Z. Sun, "A Survey on Neural Question Generation: Models, Datasets and Applications," Information Fusion, vol. 110, pp. 102015, 2024.

[9] J. Keller, P. Schaer, B. Engelmann, H. Kroll, F. Haak and C. K. Kreutz, "CIR at TREC 2024 RAG: Task 2 – Augmented Generation with Diversified Segments and Knowledge Adaptation," in Proceedings of the Text REtrieval Conference (TREC 2024), National Institute of Standards and Technology (NIST), 2024. [Online]. Available: https://trec.nist.gov/pubs/trec33/papers/irgroup.RAG.pdf

[10] X. Fu, N. S. Bedi, P. Acharya and N. Kando, "NII@TREC IKAT 2024: LLM-Based Pipelines for Personalized Conversational Information Seeking," in Proceedings of the Text REtrieval Conference (TREC 2024), National Institute of Standards and Technology (NIST), 2024. [Online]. Available: https://trec.nist.gov/pubs/trec33/papers/NII.IKAT.pdf