LLM-Based Question Generation and Retrieval-Augmented Reporting for News Credibility

Georgios Arampatzis[*]
Ioannis Maslaris
Avi Arampatzis

Democritus University of Thrace, Dept. of Electrical and Computer Engr., Xanthi, Greece
{geoaramp,imaslari,avi}@ee.duth.gr

Abstract

This paper presents the participation of the DUTH team in both tasks of the TREC 2025 DRAGUN (Detection, Retrieval, and Augmented Generation for Understanding News) Track. The track addresses the challenge of misinformation and biased narratives in digital news through two complementary tasks: **Question Generation (Task 1)** and **Report Generation (Task 2)**. Task 1 focuses on generating investigative questions that assist readers in assessing news credibility, while Task 2 evaluates systems' ability to retrieve evidence and generate grounded, well-attributed reports.

Our approach employs recent open-weight, instruction-tuned large language models (LLMs), including `Qwen2.5`, `Yi-1.5`, and `Mistral-7B`, combined with prompt engineering, semantic filtering, and retrieval-grounded generation pipelines. All systems were implemented locally using the `transformers` and `accelerate` libraries, without external fine-tuning or API access, ensuring full reproducibility and controlled model comparison.

Experimental results show that mid-sized instruction-tuned models, most notably `Mistral-7B-Instruct`, achieve the strongest rubric coverage among the DUTH submissions in the Question Generation task. In the Report Generation task, all evaluated systems exhibit very low contradiction rates, indicating robust factual grounding, but achieve limited rubric coverage under strict retrieval and attribution constraints. Overall, these findings suggest that prompt design, question specificity, and retrieval quality play a more decisive role than raw model scale in supporting explainable and evidence-based news trustworthiness assessment.

Keywords: Retrieval-Augmented Generation, News Credibility, Question Generation, Report Generation, Instruction-Tuned Language Models

1 Introduction

The increasing prevalence of misinformation and biased narratives in online news poses significant challenges for readers seeking to evaluate the trustworthiness of information sources. Traditional fact-checking approaches often attempt to verify claims directly, yet they can be time-consuming, limited in scope, and may inadvertently reinforce polarization. Recent research emphasizes the importance of *lateral reading*—encouraging readers to ask critical questions and consult multiple sources before forming judgments about credibility [1].

The TREC 2025 DRAGUN (Detection, Retrieval, and Augmented Generation for Understanding News) track continues this line of inquiry by focusing on systems that assist readers in assessing news trustworthiness through generation and retrieval tasks. In Task 1 (Question Generation), participants are asked to produce investigative questions that a critical reader should consider when evaluating an article's reliability, potential bias, or missing context. Such questions should be specific, concise, and guide further exploration—examining the author's expertise, evidence quality, or contrasting viewpoints from other outlets.

Our team, **DUTH**, participated in Task 1 by leveraging recent instruction-tuned large language models (LLMs) to automatically generate such questions. We explored the effectiveness of mid-sized open-source LLMs in producing high-quality, diverse, and well-focused investigative questions without manual intervention. Our approach emphasizes prompt design, lightweight filtering, and rule-based postprocessing to maintain compliance with task constraints ($\leq$ 300 characters, one question per line), following recent trends in LLM-driven news reasoning systems [2, 3]. Our participation builds upon prior research by the DUTH team in multilingual affective analysis, simplification, hallucination detection, and creative language generation [4–7].

[*] Corresponding author.

These earlier studies established modular pipelines for transformer-based generation and evaluation, which we extend here to the domain of news credibility assessment. The remainder of this paper is organized as follows. Section 2 reviews prior research on question generation, retrieval-augmented generation, and computational approaches to news credibility assessment. Section 3 presents the datasets, models, and methodological framework employed for both DRAGUN 2025 tasks. Section 4 reports the official evaluation outcomes for Task 1 (Question Generation) and Task 2 (Report Generation), followed by a detailed analysis of model performance and rubric-based scores. Finally, Section 5 summarizes the key findings and discusses directions for future research.

2 Related Work

Question generation has long been studied in natural language processing as a means of evaluating comprehension and guiding inquiry. Early neural models for question generation [8, 9] focused on factual question creation from educational or QA datasets such as SQuAD. With the advent of instruction-tuned LLMs [10, 11], zero-shot and few-shot generation of targeted questions has become increasingly feasible, even for higher-level reasoning and source evaluation tasks.

In the context of news credibility, the TREC 2024 Lateral Reading Track [12] and related misinformation detection efforts [13] emphasized that support systems should assist readers rather than deliver binary “true/false” labels. Instead of verifying facts, they aim to prompt critical inquiry—for instance, by suggesting investigative questions or surfacing evidence from multiple viewpoints. The DRAGUN 2025 Track extends this paradigm by linking question generation (Task 1) with retrieval-augmented report generation (Task 2), reflecting broader advances in Retrieval-Augmented Generation (RAG) [14, 15].

Our approach follows this direction by evaluating how modern LLMs can serve as cognitive scaffolds for critical news reading—prompting readers to reflect on bias, sources, and context. In particular, we examine how prompt structure and lightweight postprocessing choices relate to the specificity, diversity, and investigative focus of the generated questions.

3 System Description

3.1 Data Resources

The TREC 2025 DRAGUN Track [16] provides two principal datasets to all participating systems: (a) a large-scale segmented web corpus used for grounding and retrieval, and (b) a curated collection of thirty target news articles that serve as the primary input for both question and report generation tasks. This section describes the data resources used in each task.

3.2 Task 1: Question Generation

3.2.1 Dataset

For **Task 1 (Question Generation)**, participants were provided with a file named `trec-2025-dragun-topics.jsonl`, which contains **30 curated news articles** selected by the track organizers. Each article focuses on a distinct real-world topic drawn from a variety of media domains, including *politics, health, technology, culture,* and *science*. The dataset was designed to capture realistic variation in credibility, source framing, and evidential depth.

Each entry in the JSONL file represents one target article and includes the following fields:

- `docid`: A unique identifier linking the article to its corresponding document in the MS MARCO V2.1 [17] collection.
- `url`: The original URL of the source article.
- `title`: The headline of the article.
- `headings`: Section-level headings, if present.
- `body`: The full text of the article body, excluding advertisements and user comments.

Systems are required to generate **ten critical and investigative questions** per article, each aimed at assessing the trustworthiness, evidential grounding, and overall perspective of the source. Each question must be concise (no longer than 300 characters), clearly specific to the article's content, and submitted in a UTF-8 encoded, tab-separated file containing five fields, as defined in the official *DRAGUN* submission format. No explicit document retrieval is required for this task; however, systems may implicitly rely on background knowledge encoded during model pretraining, including knowledge derived from the MS MARCO V2.1 corpus, to support question formulation and cross-source reasoning.

3.2.2 Models

For the DRAGUN Question Generation task, we conducted a systematic evaluation of recent open-weight, instruction-tuned large language models (LLMs) that embody diverse architectural and alignment paradigms. These models were selected to capture a representative spectrum of current transformer-based generation capabilities, with emphasis on instruction following, factual grounding, and multilingual generalization—core requirements for credibility-oriented text generation.

- **Qwen2 Family** [18]: A multilingual transformer-based family of open large language models (7B, 14B, and 72B parameters) developed by Alibaba Cloud. Qwen2 models are trained on large-scale multilingual corpora and alignment data, with an emphasis on instruction following and general-purpose language understanding. They provide strong performance across a range of knowledge-intensive and reasoning-oriented benchmarks, making them suitable for credibility-oriented question generation tasks.

- **Yi 1.5–9B** [19]: An open multilingual LLM introduced by 01.AI, trained on a broad and culturally heterogeneous corpus spanning more than thirty languages. Yi 1.5 emphasizes long-context understanding, discourse coherence, and robustness to stylistic or linguistic variation—attributes essential for generating nuanced, bias-aware investigative questions across diverse media domains.
- **Mistral 7B** [20]: A compact yet competitive dense transformer model produced by Mistral AI. Although moderate in scale, it benefits from efficient attention mechanisms and high training quality, enabling stable and coherent instruction following. Owing to its strong performance–efficiency trade-off, it serves as a reliable baseline for evaluating zero-shot question generation under computational constraints.

All models were executed locally using PyTorch and the `transformers` framework [21], without any external fine-tuning, parameter updates, or API access. This setup ensures full transparency and reproducibility, isolating the intrinsic capabilities of each LLM in a zero-shot configuration. Model outputs were generated under identical decoding parameters, enabling a controlled comparative analysis of instruction adherence, linguistic diversity, and factual grounding across architectures and scales.

3.2.3 Methodology

Generation Phase. Each model listed in Section 3.2.2 was accessed via the `transformers` library [21] and executed locally using PyTorch. For every news topic provided by the organizers [16], the system generates approximately thirty question candidates from a fixed instruction prompt enforcing task constraints. The decoding configuration is standardized across all runs (temperature = 0.4, top-p = 0.85, `do_sample`=True, `no_repeat_ngram_size`=4, repetition_penalty = 1.1, seed = 42) to guarantee reproducibility.

Filtering and Ranking. Generated questions are automatically cleaned to remove duplicates, malformed text, or multi-sentence outputs. Semantic filtering relies on TF–IDF cosine similarity combined with Maximal Marginal Relevance (MMR) [22], implemented using the `scikit-learn` library [23]. This mechanism promotes coverage of distinct information facets while maintaining topic relevance, yielding a final set of ten representative questions per article.

Validation. All experimental runs are validated through the official TREC DRAGUN Evalbase interface, confirming compliance with the Task 1 specification and reproducibility criteria. No human feedback, external retrieval, or web-based augmentation is used at any stage.

3.2.4 Evaluation Measures

The evaluation of the **Question Generation** task in the TREC 2025 DRAGUN Track is based on assessor-constructed rubrics and alignment scoring between participant questions and assessor questions [16]. Each of the 30 news topics is independently reviewed by one primary and two secondary NIST assessors. Assessors conduct open-web research to determine what questions a reader should ask to evaluate the trustworthiness of each article. Their findings are merged into a single rubric per topic containing short question–answer pairs annotated with importance levels:

- **Have to Know (4 points):** Core, critical questions essential for judging the article's trustworthiness.
- **Good to Know (2 points):** Important contextual questions that enhance reader confidence but are not decisive.
- **Nice to Know (1 point):** Peripheral background questions offering additional context.

For each assessor question, two large language models, *Qwen3-Embedding-8B* and *Qwen3-Reranker-8B*, are used to automatically retrieve the most semantically similar participant-generated question. Human assessors then judge the selected question pair and assign one of four similarity labels:

- **Very Similar (1 point):** Equivalent information content despite wording differences.
- **Similar (0.5 points):** Overlapping but not identical informational scope.
- **Different (0 points):** Distinct informational focus with minor overlap.
- **Very Different (0 points):** Unrelated or dissimilar questions.

The final score for a run and topic is computed as follows. For each rubric question, if at least one submitted question is judged *Very Similar* or *Similar*, the system receives a reward equal to the product of the question's importance weight (4, 2, or 1) and the highest similarity label (1 or 0.5). The topic-level score is obtained by averaging these weighted similarities across all rubric questions. The official evaluation outputs provide both per-topic scores and aggregate statistics across submitted runs.

This design yields an interpretable, human-grounded **evaluation measure** that jointly captures question relevance, coverage, and alignment with human critical-thinking objectives, rather than relying on surface-level similarity alone. Similar rubric-based evaluation paradigms have previously been used in question answering and explainability benchmarking [24].

3.3 Task 2: Report Generation

3.3.1 Dataset

For **Task 2 (Report Generation)**, the DRAGUN Track reuses the *MS MARCO V2.1 (Segmented)* collection [17], a large-scale open-domain web corpus originally developed by Microsoft for machine reading comprehension and passage retrieval. This dataset comprises approximately **114 million text segments** derived from **11 million web documents**, with each segment representing a semantically coherent portion of a webpage from which boilerplate content has been removed.

The MS MARCO segments serve as the **retrieval base** for Task 2, enabling systems to ground their generated reports in authentic web content. Given a target article (topic) from the Task 1 dataset, systems retrieve supporting or contrasting evidence from the MS MARCO corpus and synthesize a multi-source, evidence-based report. This setup evaluates a system's ability to integrate retrieval, factual grounding, and generation in open-domain news contexts, providing a realistic benchmark for automated journalism and media literacy support.

3.3.2 Models

For the **Report Generation** task, we employ the same families of instruction-tuned large language models (LLMs) used in Task 1, extending their application to multi-source, retrieval-grounded summarization. The selection spans a range of model architectures, parameter scales, and instruction-tuning strategies, enabling an empirical comparison of factual grounding behavior across different configurations. All models are executed locally—without fine-tuning or reliance on external APIs—using the `transformers` and `accelerate` frameworks [21].

- **Qwen2.5 Series** [25]: Instruction-tuned large language models spanning 7B–72B parameters, designed for general-purpose reasoning and generation across multilingual and retrieval-augmented settings.
- **Yi 1.5–9B** [19]: An open multilingual foundation model by 01.AI, optimized for discourse coherence and multiturn dialogue. Its cross-lingual representation strength supports nuanced reasoning across heterogeneous news domains.
- **Zephyr-7B-Beta**: An alignment-tuned open model released by HuggingFace H4, trained for helpful and faithful instruction following. It serves as an efficiency-oriented reference system for evaluating grounding consistency at smaller scales.

Each model is coupled with a **BM25 retriever** implemented via `Pyserini/Anserini` over the *MS MARCO V2.1 (Segmented)* corpus [17]. For each topic, up to 40 candidate segments are retrieved, with 8–18 evidence passages retained after deduplication and length filtering. Reports are generated with deterministic decoding (temperature = 0.2, top-p = 0.9) and rigorously post-processed to ensure correct JSON structure, citation validity, and topic alignment. No external web sources (e.g., Bing, ChatGPT, or Perplexity) are accessed, guaranteeing that all system outputs remain verifiably grounded in the official TREC 2025 DRAGUN resources [16].

3.3.3 Methodology

The **Report Generation** pipeline is designed as a reproducible, three-stage framework encompassing evidence retrieval, factual grounding, and controlled report synthesis. Its objective is to produce concise, well-attributed trustworthiness summaries fully supported by verifiable text segments.

Retrieval and Evidence Selection. Document retrieval employs a BM25 framework implemented in `Pyserini/Anserini` over the *MS MARCO V2.1 (Segmented)* corpus [17]. For each target article, the system retrieves up to $k = 40$ candidate passages, retaining 8–18 top-ranked segments after deduplication and length filtering. These segments constitute the sole evidence base for report generation, ensuring complete provenance and transparency.

Grounded Report Synthesis. Each report is produced by a single forward pass of an instruction-tuned LLM from the models listed above. Prompts explicitly instruct the model to generate a 250-word analytical summary in 3–5 sentences, each citing up to three MS MARCO document identifiers (`docid`) as direct evidence. If insufficient support is found, the model is required to omit rather than infer unsupported claims, thereby reducing hallucination risk. Decoding remains deterministic across all runs to preserve comparability and reproducibility.

Post-Processing and Validation. Outputs undergo automatic validation using the official DRAGUN utility (`validate.py`), which verifies JSON syntax, citation structure, and topic consistency. Only submissions that pass all checks are uploaded to the TREC evaluation platform.

Design Rationale. The pipeline emphasizes transparency, factual attribution, and reproducibility— principles increasingly recognized as essential in generative information retrieval. By explicitly separating retrieval, grounding, and synthesis, our framework aligns with recent advances in explainable news summarization and factually grounded language generation [2, 3]. This modular design facilitates qualitative inspection of retrieval fidelity, model behavior, and evidence utilization across different model scales.

3.3.4 Evaluation Measures

The evaluation for the **Report Generation** task is conducted manually by NIST assessors following the official DRAGUN assessing protocol [16]. The process employs rubric-based scoring to capture the factual coverage, accuracy, and reliability of each submitted report, in line with prior work on structured fact evaluation and news credibility analysis [1, 2].

Rubric Construction. For each of the 30 target news articles, assessors independently formulate a set of **rubric questions** and **reference answers** ("answer nuggets") after reviewing the article and conducting independent web verification. Each question is assigned an importance weight according to its contribution to evaluating the article's trustworthiness:

- **Have to Know (4 points)** – essential information for determining factual accuracy.
- **Good to Know (2 points)** – relevant contextual information that supports judgment.
- **Nice to Know (1 point)** – peripheral but useful background details.

Questions for which no supporting evidence can be retrieved from the *MS MARCO V2.1 (Segmented)* collection [17] are removed from the rubric prior to scoring.

Scoring Framework. Each generated report is manually aligned against the rubric nuggets and annotated using one of four relevance labels:

- **Supports** – provides a correct, consistent, and evidence-backed answer.
- **Partial** – captures only part of the correct information.
- **Contradicts** – contains statements conflicting with reference answers.
- **None** – provides no relevant information.

Score Computation. For each topic, two complementary **evaluation measures** are computed:

- **Supportive Score** (0–1, higher is better): measures factual alignment with rubric nuggets. A *Supports* label contributes 1.0 and *Partial* 0.5, weighted by each question's importance level. The total is normalized by the maximum possible weighted score.
- **Contradictory Score** (0–1, lower is better): penalizes contradictory or misleading content, where each *Contradicts* label contributes 1.0.

Interpretation. The **Supportive Score** reflects how effectively a system's report addresses assessor-defined questions, while the **Contradictory Score** captures the frequency of factual inconsistencies. Together, these **evaluation measures** assess a system's ability to perform grounded, evidence-based reporting and align with principles of lateral reading and cross-source verification [1]. This evaluation approach builds upon earlier work in question-based trustworthiness assessment [2, 3], providing a consistent framework for assessing generative systems in journalism-oriented information retrieval.

3.4 Implementation and Environment

All experiments for both tasks are implemented in Python 3.10 using the Hugging Face `Transformers` and `Accelerate` frameworks [21]. Execution is carried out on a high-performance GPU workstation at the Democritus University of Thrace (DUTH), equipped with an NVIDIA RTX A6000 (48 GB VRAM). Model checkpoints and tokenizers are retrieved from the Hugging Face Hub under authenticated access, with local caching managed through the `HF_HOME` environment variable to ensure efficient reuse and full reproducibility.

All experimental runs follow a unified pipeline based on the official DRAGUN starter kit. The framework automatically loads the selected instruction-tuned model (`Qwen2.5-7B-Instruct`, `Mistral-7B-Instruct-v0.3`, or `Yi-1.5-9B-Chat`) and produces outputs compliant with the Task 1 and Task 2 submission formats. Quantized inference is performed using the `BitsAndBytesConfig` API with 4-bit precision to enable memory-efficient deployment [26]. Consistent random seeds, decoding parameters, and generation settings are maintained across runs to guarantee replicability. All generated files are validated locally using the official `trec_dragun_validate.py` utility before submission to Evalbase.

4 Results

4.1 Task 1: Question Generation

Table 1 reports the results of the DUTH (garamp) submissions for Task 1 (Question Generation) across different instruction-tuned large language models. All runs follow the same experimental protocol, using zero-shot inference, fixed decoding parameters, and fully local execution without external APIs or fine-tuning. This setup enables a direct and controlled comparison among the evaluated models.

As shown in Table 1, the Mistral-7B model achieves the highest mean score (0.199) among the DUTH runs. The Yi-1.5-9B (0.193) and Qwen2.5-72B (0.177) models follow with slightly lower scores, while the remaining configurations exhibit reduced rubric coverage. This

Table 1: Task 1 (Question Generation) results for the DUTH team, as reported in the official TREC 2025 DRAGUN evaluation [27]. Average scores across 30 topics.

Team	Run ID	Score
DUTH	mistral_7b	**0.199**
DUTH	yi15_9b	0.193
DUTH	qwen25_72b	0.177
DUTH	qwen25_7b_imp	0.158
DUTH	qwen25_14b	0.149

ranking indicates that, within the examined pipeline, mid-sized instruction-tuned models achieve the strongest performance among the DUTH submissions for investigative question generation.

A comparison across the DUTH submissions suggests that increased model size does not necessarily lead to improved performance on Task 1. Despite their larger parameter counts, models such as Qwen2.5-72B do not outperform the smaller Mistral-7B under the same experimental conditions. This observation is consistent with the rubric-based evaluation framework, which prioritizes rubric coverage and semantic alignment with assessor-defined investigative criteria over raw model capacity.

Overall, the results in Table 1 demonstrate that a reproducible, lightweight pipeline—combining prompt-based generation with semantic filtering and selection—can yield stable results for credibility-oriented question generation without reliance on large-scale models or external optimization.

4.2 Task 2: Report Generation

Table 2 reports the average supportive and contradictory scores across 30 topics for Task 2 (Report Generation). The supportive score measures the extent to which a generated report provides correct and rubric-aligned information, while the contradictory score captures the presence of statements that conflict with assessor-defined answer nuggets. Higher supportive and lower contradictory scores indicate better performance.

As shown in Table 2, all evaluated systems achieve relatively low absolute supportive scores. This outcome is expected given the constraints of the task, including the fixed 250-word limit, the requirement to cite evidence exclusively from the MS MARCO V2.1 Segmented Corpus, and the fact that assessor rubrics were constructed using open-web research that may extend beyond the available retrieval corpus.

Focusing on the DUTH (garamp) submissions, all reported runs exhibit identical average supportive scores (0.005) and very low contradictory scores (0.002). This indicates that, within the evaluated submissions, the generated reports rarely contradicted assessor rubrics but also covered only a limited portion of the rubric-defined answer nuggets. The consistency of these scores across different instruction-tuned models suggests that, under the examined pipeline, model choice had a limited impact on Task 2 performance.

Table 2: Task 2 (Report Generation) results for the DUTH (garamp) team, as reported in the official TREC 2025 DRAGUN evaluation [27]. Average supportive and contradictory scores across 30 topics. Higher supportive and lower contradictory scores indicate better performance.

Team	Run ID	Supportive	Contradictory
DUTH	dragun_t2_q7b	0.005	0.002
DUTH	qwen25_14b_r4	0.005	0.002
DUTH	qwen25_3b_t2	0.005	0.002
DUTH	yi9b_t2_v1	0.005	0.002
DUTH	zephyr7b_t2	0.005	0.002

Overall, the results in Table 2 highlight the difficulty of the Report Generation task under strict grounding and attribution constraints. They further suggest that achieving higher rubric coverage may depend more strongly on retrieval effectiveness, evidence selection, and report structuring strategies than on the underlying language model alone.

5 Conclusion and Future Work

This work presented the participation of the DUTH team in the TREC 2025 DRAGUN Track, covering both Task 1 (Question Generation) and Task 2 (Report Generation). Through systematic experimentation with open-weight, instruction-tuned large language models (LLMs), we examined how model selection, prompt design, and grounding constraints affect the ability of generative systems to support news trustworthiness assessment under fully reproducible settings.

For Task 1, our results show that mid-sized instruction-tuned models, most notably **Mistral-7B-Instruct**, achieve the strongest performance among the DUTH submissions. The observed ranking across models indicates that increased parameter count does not necessarily lead to improved rubric coverage in investigative question generation. These findings suggest that alignment with assessor-defined criteria, prompt adherence, and question specificity play a more influential role than raw model scale within a zero-shot, locally executed pipeline.

For Task 2, all DUTH submissions yield uniformly low supportive scores and very low contradictory scores. This outcome reflects the strict grounding requirements of the task, the limited evidence available within the MS MARCO V2.1 corpus [28], and the use of a shared retrieval and generation pipeline across models. While the generated reports rarely contradict assessor rubrics, they cover only a small portion of the rubric-defined answer nuggets, indicating that report completeness remains a significant challenge under constrained retrieval settings.

Overall, the results highlight that transparent, lightweight pipelines can produce stable and reproducible behavior across both tasks, but also underscore the importance of improved retrieval strategies and report structuring to increase rubric coverage, particularly for Task 2. These observations are consistent with prior work on retrieval-augmented generation and grounded text production in high-stakes information settings [16, 21].

For future work, we plan to investigate enhanced retrieval pipelines that combine sparse and dense retrievers, as well as evidence-aware prompt designs that more explicitly guide report generation toward answering high-importance rubric items. We also aim to explore uncertainty-aware and explanation-focused generation strategies, with the goal of improving coverage and interpretability while maintaining strict factual grounding in support of human-centered media literacy.

References

[1] Sam Wineburg and Sarah McGrew. Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, 121(11):1–40, 2019.

[2] X. Liu, C. Li, and M. Sun. Assessing news trustworthiness via multi-view question generation. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

[3] Z. Hu, Y. Lin, and X. Ma. Prompt-based fact checking with large language models. *arXiv preprint arXiv:2401.11254*, 2024.

[4] Georgios Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. Duth at semeval-2023 task 9: An ensemble approach for twitter intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1225–1230, 2023. URL `https://aclanthology.org/2023.semeval-1.170/`.

[5] Georgios Arampatzis and Avi Arampatzis. Duth at clef 2025 simpletext track: Tackling scientific text simplification and hallucination detection. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[6] Georgios Arampatzis et al. Duth at exist 2025: Multilingual sexism detection with soft labels and transformers. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[7] Georgios Arampatzis and Avi Arampatzis. Duth at clef joker 2025 tasks 2 and 3: Translating puns and proper names with neural approaches. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[8] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *ACL 2017*, pages 1342–1352, 2017.

[9] Wangchunshu Zhou et al. Generating question-answer hierarchies. In *ACL 2019*, pages 5632–5642, 2019.

[10] Long Ouyang et al. Training language models to follow instructions with human feedback. In *NeurIPS 2022*, 2022.

[11] Hugo Touvron et al. Llama: Open and efficient foundation language models. In *ICLR 2023*, 2023.

[12] Shuyang Gao, Cheng Li, Ellen Voorhees, and Ian Soboroff. Overview of the trec 2024 lateral reading track. In *TREC 2024*, 2024.

[13] Alberto Barrón-Cedeño et al. Overview of the clef-2023 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *CLEF 2023 Working Notes*, 2023.

[14] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS 2020*, 2020.

[15] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021*, pages 874–880, 2021.

[16] TREC DRAGUN Organizers. Trec dragun track 2025: Detection, retrieval, and generation for understanding news (dragun). `https://trec-dragun.github.io/`, 2025. TREC 2025 Track Overview.

[17] Tri Nguyen et al. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, 2016.

[18] Yitao Bai et al. Qwen2: A family of open large language models from alibaba cloud. `https://huggingface.co/Qwen`, 2024. Versions: Qwen2-7B, Qwen2-14B, Qwen2-72B.

[19] R. Young et al. Yi: Open foundation and chat models by 01.ai. `https://huggingface.co/01-ai/Yi-1.5-9B`, 2024. Yi-1.5 series, including 9B parameter model.

[20] Albert Jiang et al. Mistral 7b: Open-weight efficient language model. `https://mistral.ai/news/announcing-mistral-7b/`, 2023. Dense 7B parameter model released by Mistral AI.

[21] Thomas Wolf et al. Transformers: State-of-the-art natural language processing. In *EMNLP 2020: System Demonstrations*, pages 38–45, 2020.

[22] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 335–336, 1998.

[23] Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Jeffrey Dalton et al. Cast 2021: The conversational assistance track overview. In *Proceedings of TREC 2021*. NIST, 2021.

[25] Qwen Team. Qwen2.5: Bridging multilingual and multimodal llms. Alibaba Cloud, 2024. `https://huggingface.co/Qwen`.

[26] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. ML Research Press, 2023. arXiv:2305.14314.

[27] 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 *Proceedings of the Thirty-Fourth Text REtrieval Conference (TREC 2025)*, Gaithersburg, MD, USA, 2025. National Institute of Standards and Technology (NIST).

[28] A. Dalmia, S. MacAvaney, et al. Ms marco v2.1: A segmented large-scale open-domain corpus for passage retrieval. *Microsoft Research*, 2021.