Discourse-guided Lateral Reading for Automated News Trustworthiness Assessment

Oliwia Majtyka and Ting Liu
Department of Computer Science, Siena University, Loudonville, New York, USA

o20majt@siena.edu
tliu@siena.edu

Abstract: The rapid proliferation of misinformation across digital news platforms has created an urgent demand for assistive technologies that empower users to evaluate the credibility of online content. Lateral reading, the investigative practice of verifying claims via external sources, is widely recognized as a highly effective strategy for assessing trustworthiness. However, existing automated approaches to question generation (QG) often falter in this context, producing generic or holistic inquiries that fail to target the specific claims, evidence, and named entities within an article, thereby limiting their utility for rigorous real-world fact-checking. In this paper, we present a two-stage pipeline for discourse-guided lateral reading question generation, specifically developed for the TREC 2026 DRAGUN (Detection, Retrieval, and Augmented Generation for Understanding News) Track. Our approach addresses a critical limitation of document-level QG by shifting the focus to fine-grained discourse segments. In the first stage, we utilize a few-shot prompted large language model (LLM) to decompose news articles into functional units, such as claims, evidence, fact descriptions, and supporting context, grounded in the principles of Rhetorical Structure Theory (RST). By identifying the hierarchical and relational roles of these segments, the second stage of our system generates targeted investigative questions tailored to the specific discourse type of each unit. Preliminary evaluation of our discourse-guided pipeline demonstrates a significant improvement in question specificity and relevance when compared to a traditional single-stage baseline. To provide a more formal and standardized validation of these results, we are currently employing the TREC AutoJudge framework. This meta-track evaluation allows us to score our system against importance-weighted rubrics and expert-derived investigative priorities, ensuring the generated questions facilitate objective verification. Our hypothesis is that this structural decomposition significantly reduces low-quality, "summary-style" questions in favor of precise inquiries that ground each prompt in a verifiable claim. These findings suggest that discourse-guided segmentation is a scalable and effective strategy for enhancing the precision of automated fact-checking, with broader implications for retrieval-augmented generation (RAG), media literacy tools, and the development of more transparent, verifiable AI systems.

Keywords: Lateral reading, Misinformation and disinformation detection, Question generation, Discourse segmentation, Few-shot prompting, News trustworthiness

1. Introduction

The problem of news trustworthiness has become one of the defining challenges of the information age. Readers encountering online news articles must determine whether a source is credible, whether the claims made are accurate, and whether the framing of a story is fair. Traditional fact-checking is labour-intensive and cannot scale to the volume of content produced daily. Automated support for news credibility evaluation is therefore both an important research direction and a pressing practical need.

Lateral reading offers one promising approach. Developed and validated in the context of professional fact-checking, lateral reading involves consulting external sources to investigate a publisher, author, or claim rather than reading the article itself in depth (Wineburg and McGrew, 2017). Studies have shown that experienced fact-checkers rely heavily on this strategy to rapidly assess source credibility, while novice readers tend to read vertically, spending time within the document rather than checking it against external evidence (Breakstone et al., 2021). The key insight is that the most useful questions for evaluating a news article are often not about its internal content but about who wrote it, who published it, and whether the named claims, experts, and institutions it cites have a verifiable track record.

Automating lateral reading presents a distinct challenge from related tasks such as question answering or summarisation. A lateral reading question must be specific enough to yield useful search results, grounded in a named entity or claim from the article, answerable by someone who has never read the article, and focused on external verification rather than comprehension. Generic question generation systems tend to produce questions that could apply to any article, failing to capture the article-specific details that matter most for trustworthiness assessment.

This paper presents a two-stage pipeline that addresses these requirements through discourse-guided segmentation followed by segment-conditioned question generation with human-scored few-shot examples. The initial single-stage pipeline was submitted to the TREC 2025 Lateral Reading track; the improved two-stage

architecture was developed as an extension of our TREC 2025 DRAGUN submission. This paper documents both the 2025 baseline and four pipeline configurations demonstrating progressive gains from discourse architecture improvements and iterative selection strategy refinements.

Our results show that discourse-guided generation produces more evaluator-aligned questions than unstructured baselines, with each configuration demonstrating consistent gains.

2. Related Work

Lateral reading as a fact-checking strategy was established by Wineburg and McGrew (2017) and empirically validated by Breakstone et al. (2021). The TREC 2024 Lateral Reading track (Zhang et al., 2024) formalised automated lateral reading as an NLP task scored using NDCG, and the TREC 2025 DRAGUN track extends this framework with increasing focus on discourse structure. This paper presents our extended system developed beyond the submission. The highest-scoring TREC 2024 system, h2oloo (University of Waterloo), used a stepwise GPT-4o decomposition strategy to achieve NDCG@10 of 0.735 (h2oloo, 2024). Our work builds on this insight by making that decomposition principled and explicit through discourse-guided rhetorical segmentation. In our own TREC 2025 DRAGUN submission (Cheverton, Majtyka and Liu, 2025), we explored query expansion and LLM-augmented generation for both question and report generation, which form the baseline from which the improvements described in this paper were developed.

The theoretical basis for our segmentation stage is Rhetorical Structure Theory (Mann and Thompson, 1988), with annotation standards drawn from the RST Discourse Treebank (Carlson and Marcu, 2001). Rather than implementing full RST parsing, we define five task-specific segment categories (Claim, Evidence, Fact Description, Event Description, Supporting Comments) informed by RST principles but optimised for news trustworthiness assessment. Full RST parsers were considered and rejected for three reasons: they require fine-tuned models trained on general news text rather than trustworthiness-oriented annotation; their relation labels such as elaboration and contrast do not map cleanly to investigation priorities; and they struggle with entity-dense reporting text without fine-tuning (Wang et al., 2024). Few-shot prompted segmentation was selected instead because it allows task-specific category definitions without additional training data. Neural RST parsing approaches such as DMRST (Liu et al., 2021) demonstrate that segmentation and relation classification can be unified, while recent evaluations show LLMs struggle with complex RST relation types without fine-tuning (Wang et al., 2024), motivating our few-shot prompting approach over zero-shot alternatives. Most directly related is Desai and Moldovan (2021), who apply segment classification to condition question generation. Our work adapts this from reading comprehension to lateral reading trustworthiness assessment.

Few-shot prompting (Brown et al., 2020) underpins our strategy of teaching quality criteria through scored positive and negative examples rather than explicit rules.

Taken together, these works leave a clear gap this paper addresses, motivating the central research question: can principled discourse segmentation produce lateral reading questions that better match expert investigative priorities compared to unstructured or implicit decomposition? Prior sentence-level QG systems produce questions without grounding in the document's argumentative structure. Automated fact-checking pipelines focus on classification and verdict prediction rather than generating actionable investigative questions. Prior lateral reading systems, including our own TREC 2025 submission, decompose articles implicitly through prompting without typed rhetorical roles, limiting reproducibility and segment-conditioned generation. This paper addresses all three gaps by defining a structured mapping from discourse roles to investigation priorities and empirically testing whether principled segmentation produces questions that better match what human assessors consider most important for trustworthiness evaluation.

3. Methodology

We present a discourse-to-investigation transformation framework that converts a raw news article into a ranked list of at most ten lateral reading questions. The first stage assigns rhetorical roles to segments, treating each as an epistemic object whose type determines what external verification is most useful. The second stage applies investigation operators over those typed segments, then applies an epistemic prioritisation policy to select the ten most valuable questions. Both stages use GPT-4o (OpenAI, 2024) with few-shot prompting. GPT-4o was selected over smaller open-weight models because the top-performing TREC 2024 system demonstrated its effectiveness for this task, and because structured discourse annotation requires strong instruction-following capabilities that smaller models did not reliably provide in preliminary testing. A two-stage design was chosen over a single-stage approach because separating segmentation from question generation allows each stage to

be tuned independently, makes the pipeline more interpretable, and enables segment-type conditioning that a monolithic prompt cannot easily achieve. Figure 1 illustrates the architecture.



Figure 1: Architecture of the two-stage discourse-guided lateral reading pipeline

Table 1: Prompt configuration parameters for each pipeline stage

Stage	Model	Few-shot examples	Temperature	N runs
Segmentation	GPT-4o	4 annotated articles (health, politics, environment, multi-party)	0.3	1
Question generation	GPT-4o	3 scored examples (score 4 ideal vs score -1 bad, with explanations)	0.7	1

3.1 Preprocessing

Each article is preprocessed before Stage 1. Boilerplate lines are removed using a regular expression filter covering subscription prompts, social sharing labels, advertisement markers, copyright notices, and social media platform references. Lines under 20 characters without sentence-final punctuation are discarded. Heading noise labels are filtered, retaining only the first three clean headings. The body is then truncated at 6,000 characters at a sentence boundary, with a [truncated] marker appended. This limit was chosen empirically: 6,000 characters captures the argumentative core of most news articles, which typically front-load key claims and evidence, while remaining within GPT-4o's practical context window. Longer thresholds produced diminishing segmentation quality gains while substantially increasing API costs.

3.2 Stage 1: Discourse Segmentation

The first stage decomposes the article into its core rhetorical units using a prompted segmentation model. The model is instructed to identify four to eight passages that represent the article's argumentative skeleton, prioritising: the central claim the author wants the reader to accept; named sources, institutions, studies, or statistics cited as support; explicit assertions of fact that a reader would need to verify independently; and contextual framing that shapes interpretation. Each identified passage is annotated across five categories: Fact Description, Event Description, Claim, Supporting Comments, and Evidence. The segmentation prompt runs at temperature 0.3 to encourage deterministic, structured output; annotation tasks benefit from low variance since the same article should consistently produce the same segment boundaries and labels.

A critical design feature is the explicit extraction of article metadata: the publisher and author names as they appear in the article. This metadata anchors author credibility and publisher reputation questions in Stage 2, two of the most valued question types in lateral reading assessments, ensuring they appear even when byline information is absent from the selected segments.

The segmentation prompt includes four annotated few-shot examples covering a range of article types: a health and nutrition article, a political controversy piece, an environmental policy story, and a multi-party blame attribution narrative. Each example demonstrates not only the correct annotation format but also the reasoning behind segment selection. To illustrate, one segment extracted from a CBS News article on the Fukushima water discharge reads: original text: the IAEA supports this plan as scientifically reasonable; segment type: Claim; annotation: a high-stakes institutional endorsement whose accuracy is independently verifiable through IAEA public statements, making it a priority segment for question generation.

3.3 Stage 2: Lateral Reading Question Generation

The second stage takes the segmented article, including all segment annotations and the extracted metadata, and generates lateral reading questions. The question generation prompt is structured around a scoring rubric drawn from TREC DRAGUN evaluation criteria, defining five quality levels from score 4 (specific, externally verifiable, tied to a named entity) down to score -1 (answered by the article itself, or relying on vague references such as "the study mentioned" or "the expert cited").

The prompt includes three scored few-shot examples presenting positive and negative question examples with explanations. A key failure mode shown by the negative examples is treating an article as a text to evaluate rather than a set of claims to verify externally, producing journalism meta-questions such as "Does the article present balanced viewpoints?" rather than targeted questions such as "What is the track record of [named institution] in [specific domain]?"

Question generation is conditioned on segment type. Claim segments prompt the model to challenge the assertion and ask whether it is supported by independent sources. Evidence segments prompt questions about the methodology and credibility of the cited source. Fact Description segments prompt verification of the specific named entity, statistic, or institution. Segments with no verifiable content generate no questions.

Beyond segment-level generation, the prompt instructs the model to consider four additional trustworthiness dimensions: publisher and author credibility (using the extracted metadata); definitional and background context; outcome and follow-up for ongoing situations or legal proceedings; and missing context that the article assumes the reader already has.

The model generates all questions in a single pass at temperature 0.7, which encourages diversity in the candidate question pool while remaining grounded in the article content. A higher temperature is appropriate here because question generation benefits from lexical variation across candidates. The model then self-selects the ten best questions from its own output using four explicit criteria stated in the prompt: specificity to a named entity or claim in the article; external searchability, meaning the question can be answered without reading the article; topical coverage across different aspects of the article; and the absence of near-duplicate questions. A final deduplication step is applied as a post-processing safeguard: questions are compared pairwise using a word-overlap ratio (intersection over minimum length), and any question sharing more than 0.75 of its words with an already-selected question is discarded.

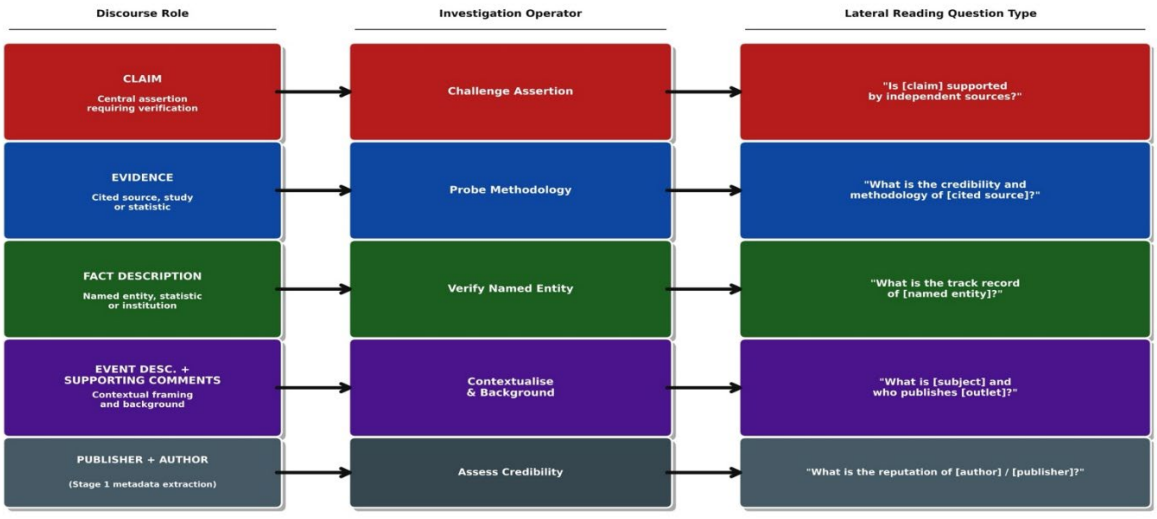

Figure 2: Discourse-to-investigation mapping showing how each rhetorical role conditions a corresponding investigation operator and lateral reading question type

3.4 Subject-aware Question Selection

Building on v3, the v3.2 configuration introduces a subject-awareness mechanism into the question selection step. The selection prompt instructs the model to first assess whether the article's central subject is widely known or relatively obscure to a general reader, then conditions selection differently based on that assessment.

For articles introducing a technology, institution, person, or event that is not widely known and is not explained within the article itself, the selection criteria prioritise a definitional question as the single most important output. The reasoning is that a reader who does not understand what BrainEx, the Blue Whale Challenge, or Gastro Obscura is cannot meaningfully evaluate any of the article's specific claims. Providing a 'What is X?' question at rank one gives that reader the foundational context needed to proceed. Specific verification questions are then selected to fill the remaining nine slots.

For articles about widely known subjects where the topic requires no special context, such as a major political figure, a well-known technology company, or a familiar cultural event, the selection criteria skip the definitional question and prioritise specific, externally verifiable claim verification questions instead. This prevents the model from generating a generic definitional question that adds no value for a reader who already understands the domain, freeing a slot for a higher-quality investigative question.

This two-path logic addresses a failure mode in v3 and v3.1: on articles about well-known subjects, those systems occasionally wasted a slot on a definitional question a general reader would not need. Subject-aware selection ensures the highest-ranked questions address the reader's actual knowledge gap before moving to specific claim verification.

4. Experimental Setup

We evaluated our pipeline on the TREC 2025 DRAGUN dataset, comprising 30 news articles from the MS MARCO V2.1 Document Corpus covering controversial topics from 2019 to 2021 across 28 diverse media sources (Zhang et al., 2025). Evaluation uses assessor-built rubrics with importance-weighted questions: Have to Know (4 points), Good to Know (2 points), and Nice to Know (1 point). Participant questions earn credit when judged Very Similar (1.0) or Similar (0.5) to a rubric question, scaled by importance weight and normalised to [0, 1]. The TREC AutoJudge framework, achieving Kendall's $\tau = 0.678$ with official human judgements (Zhang et al., 2025), provides automated scoring against the same rubrics.

We developed and compared four pipeline configurations. The baseline (v1) is our original single-stage submission to the TREC 2025 Lateral Reading track, using a simpler prompting approach without discourse segmentation. The primary contribution of this paper is the discourse-to-investigation framework realised in v3.2, the best-performing configuration, developed as an extension of our TREC 2025 DRAGUN submission. v3 established the core two-stage architecture: four-example discourse segmentation, explicit metadata extraction, segment-type-conditioned question generation with a scored few-shot rubric, and the four-dimensional trustworthiness instruction. v3.1 refined the selection criteria to encourage definitional questions for unfamiliar subjects. v3.2 formalised this into the subject-aware conditional selection mechanism described in Section 3.4, achieving the highest score across all configurations and representing the strongest configuration of our extended system.

All experiments used GPT-4o with the configuration described in Section 3. Article text was preprocessed to remove boilerplate content, including navigation elements, social sharing prompts, and copyright notices, then truncated to 6000 characters at a sentence boundary to fit within the model's practical context window for this task.

5. Results

Table 2 reports the average scores for all Task 1 (Question Generation) runs submitted to TREC 2025 DRAGUN, reproduced from the track overview paper (Zhang et al., 2025). Our submission (SCIAI Team02 Task1) achieved a score of 0.304, ranking 3rd overall among 37 runs from 10 teams, behind the Perplexity Deep Research organiser baseline (0.354) and h2oloo's GPT-5 original prompt (0.307). The top four runs are not statistically significantly different from one another, indicating that our submission is competitive with the strongest systems in the track.

Table 2: Average scores across 30 topics for Task 1 (Question Generation). Reproduced from Zhang et al. (2025)

Team / Run ID	Score	Notes
organizer-t1-perplex (Perplexity)	0.354	*Organiser baseline*
h2oloo gpt5 orig	0.307	
SCIAI Team02 Task1 (ours)	0.304	*3rd place*
h2oloo gpt5 step	0.294	
cycraft cursor-enhanced	0.264	
organizer-gpt-oss-t1	0.235	*Organiser baseline*
organizer-t1-chatgpt (ChatGPT)	0.231	*Organiser baseline*
... (30 additional runs)	0.199 – 0.081	

Table 3 shows the performance of all four pipeline configurations. The improved configurations (v3, v3.1, and v3.2) were developed after the official TREC 2025 submission deadline and were evaluated following the official TREC 2025 DRAGUN evaluation process. Our original submission (v1) achieved a score of 0.304, placing 3rd among 37 runs, and serves as the baseline for comparison.

Table 3: Performance of pipeline configurations evaluated following the official TREC 2025 DRAGUN evaluation process

Configuration	Score	Description
v1 (baseline)	0.304	Official TREC 2025 submission (3rd/37 runs)
v3	0.352	Two-stage discourse-guided pipeline
v3.1	0.384	Refined selection for unfamiliar subjects
v3.2 (best)	0.387	Subject-aware conditional selection

The primary improvement comes from the transition from the single-stage baseline (v1) to the discourse-guided architecture introduced in v3. The best configuration, v3.2, achieves a score of 0.387 compared to 0.304 for v1, a relative improvement of 27.3%. This gain reflects the cumulative effect of the two-stage discourse segmentation, explicit metadata extraction, scored few-shot question generation, and subject-aware selection strategy. The progression from v3 (0.352) to v3.1 (0.384) reflects improved handling of definitional questions for unfamiliar subjects, while the further step to v3.2 (0.387) reflects the subject-aware conditional selection mechanism. The difference between v3.1 and v3.2 is modest, and the primary driver of the overall improvement is the discourse-guided architecture rather than the selection refinements alone.

6. Discussion

Our results demonstrate that discourse structure provides a meaningful scaffold for lateral reading question generation. GPT-4o was used identically across all configurations, suggesting that architectural and prompting changes rather than model capability drive the observed improvements. By decomposing an article into its argumentative units, the pipeline grounds each question in a specific verifiable passage, reducing the tendency toward generic journalism meta-questions in favour of targeted entity-level questions a reader can search for externally.

The inclusion of scored negative examples in the question generation prompt proved particularly important. Without explicit examples of what a bad lateral reading question looks like and an explanation of why it fails, models default to producing meta-questions that evaluate the article's internal logic rather than directing the reader to external verification sources. The contrast between score 4 and score -1 examples teaches the model a distinction that is difficult to specify as a rule: the difference between "What is the IAEA's published standard for safe radiation levels in ocean discharge?" (specific, external, actionable) and "What evidence supports the claim that the water will be safe?" (effectively answered by the article itself).

The metadata extraction mechanism in Stage 1 directly addresses one of the most consistently valued question types in lateral reading assessments: who wrote and who published the article. Since byline information does not always appear in the body passages selected for segmentation, explicit extraction ensures that publisher and author credibility questions are always present in the final output.

The largest gains arise from introducing explicit discourse segmentation and segment-conditioned question generation, with the step from v1 to v3 accounting for most of the overall improvement. The subsequent selection refinements in v3.1 and v3.2 provide smaller additional gains: v3.1 scores 0.384 and v3.2 scores 0.387 under AutoJudge, a modest difference of 0.003. While v3.2 achieves the highest score among all evaluated configurations, the primary contribution of this work is the discourse-guided architecture rather than the subject-aware selection mechanism alone.

The value of v3.2's subject-awareness logic is most evident on niche or specialised articles, where the pipeline previously generated strong verification questions but missed the basic contextual question a general reader needs first. The mechanism corrects this by asking whether the reader needs to understand what the subject is before evaluating any claims. When yes, a definitional question takes a top-ranked position; when no, that slot goes to an additional verification question, better aligning the output with how an expert fact-checker would prioritise.

Together, the three improvements represent a complementary architecture: discourse segmentation grounds questions in specific verifiable content; the scored few-shot rubric distinguishes useful from generic questions; and subject-aware selection calibrates output to the reader's knowledge state. Each layer addresses a different failure mode.

Table 4 shows a qualitative comparison between v1 and v3.2 on a BuzzFeed News article about the Havana Syndrome NAS report. v1 produces vague questions about mass hysteria and general methodology; v3.2 anchors each question to named entities in the article.

Table 4: Qualitative comparison of v1 and v3.2 questions for a BuzzFeed News article on the Havana Syndrome NAS report

v1 (baseline)	v3.2 (best)
Are there any documented cases of mass hysteria leading to symptoms similar to Havana Syndrome?	What evidence does the National Academies of Sciences report provide to support the claim that microwave attacks are the most plausible explanation for Havana Syndrome?
What methodologies did the National Academies of Sciences use to evaluate the multiple theories regarding Havana Syndrome?	What is the State Department's history of involvement with investigations into Havana Syndrome?

v1 (baseline)	v3.2 (best)
What potential biases might the NAS panel have in favour of the microwave attack explanation?	Who are the external experts on microwaves and Havana Syndrome who dismiss the NAS report's conclusions, and what are their credentials?
What are the known biases of BuzzFeed News in reporting on government-related health issues?	What is Dan Vergano's track record as a journalist in reporting on scientific and health-related topics?

One limitation of the current pipeline is its reliance on GPT-4o for both stages, which introduces cost and latency constraints for large-scale deployment. Future work could explore distilled models for the segmentation stage, which is the more structured and reproducible of the two tasks. A second limitation is the fixed truncation at 6000 characters, which may cause the pipeline to miss verifiable claims appearing later in longer articles. Adaptive truncation strategies that preserve the most claim-dense passages could address this without increasing input length. Finally, while AutoJudge provides a scalable and validated approximation of human assessment, future work should confirm that the observed gains transfer under full official human evaluation to ensure they are not an artifact of automated rubric matching.

7. Conclusion

We have presented a discourse-to-investigation framework for automated lateral reading question generation grounded in the rhetorical structure of news articles. The work spans two phases: a single-stage submission to TREC 2025 and the redesigned two-stage pipeline developed as an extension of our TREC 2025 DRAGUN submission.

The contributions of this work are three conceptual advances. First, a discourse-to-investigation mapping: a principled framework that assigns rhetorical roles to article segments and derives investigation priorities from those roles, providing an explicit connection between discourse structure and external verification in the lateral reading context. Second, epistemic-role-based question conditioning: a question generation design that uses scored positive and negative examples to teach a model the distinction between questions that verify claims externally and questions that merely evaluate the article's internal logic. Third, reader-state-aware prioritisation: a subject-awareness mechanism that conditions question selection on an assessment of the reader's prior knowledge, ensuring the highest-ranked questions address the reader's actual knowledge gap rather than defaulting to a fixed question structure.

We believe that discourse-guided lateral reading automation has substantial potential as a component of broader news trustworthiness tools. The pipeline family described here offers a transparent and reproducible foundation for future research in automated media literacy support. Several directions merit investigation. These include evaluating newer frontier models such as GPT-4.1 as drop-in replacements; adaptive truncation strategies that vary the character limit based on article length; evaluation on larger article sets beyond the 30-topic TREC dataset; and extending the pipeline to surface disinformation signals alongside trustworthiness assessment. The evaluation framework established by the TREC lateral reading initiative provides a valuable benchmark for continued progress in this direction.

Acknowledgements

The authors thank Siena University for supporting this research through the Committee on Teaching and Faculty Development (COTFD) and the Center for Undergraduate Research and Creative Activity (CURCA). We also express our gratitude to the TREC DRAGUN organizers for providing the evaluation framework.

Ethics Declaration: This research did not involve human participants, personal data, or sensitive materials. Ethical clearance was not required.

AI Declaration: Claude (Anthropic) was used to assist with language editing and proofreading of this paper. All technical content, experimental design, results, and analytical conclusions are the work of the authors. The authors take full responsibility for the accuracy and integrity of the submitted paper.

References

Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M. and Saavedra, A. (2021) "Students' Civic Online Reasoning: A National Portrait", Educational Researcher, Vol 50, No. 8, pp 505-515. doi:10.3102/0013189X211017495

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. and Amodei, D. (2020) "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, Vol 33, pp 1877-1901.

Carlson, L. and Marcu, D. (2001) Discourse Tagging Reference Manual, USC Information Sciences Institute Technical Report, Los Angeles.

Cheverton, J., Majtyka, O. and Liu, T. (2025) "Intelligent News Comprehension through Query Expansion and LLM-Augmented Generation", The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025), NIST Special Publication, National Institute of Standards and Technology.

Desai, T. and Moldovan, D.I. (2021) Discourse Parsing and its Application to Question Generation, PhD Dissertation, University of Texas at Dallas.

h2oloo (2024) "h2oloo at TREC 2024 Lateral Reading Track", Proceedings of the 33rd Text REtrieval Conference, NIST.

Liu, L., Shi, W., Huang, M. and Zhu, X. (2021) "DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Relation Classification", Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 3530-3546. doi:10.18653/v1/2021.emnlp-main.278

Mann, W.C. and Thompson, S.A. (1988) "Rhetorical Structure Theory: Toward a Functional Theory of Text Organisation", Text, Vol 8, No. 3, pp 243-281. doi:10.1515/text.1.1988.8.3.243

OpenAI (2024) GPT-4o Technical Report, OpenAI, San Francisco. https://openai.com/research/gpt-4o

Wang, J., Dong, Y. and Li, J. (2024) "RST Discourse Parsing with Large Language Models", arXiv preprint arXiv:2403.05065.

Wineburg, S. and McGrew, S. (2017) "Lateral Reading: Reading Less and Learning More When Evaluating Digital Information", Stanford History Education Group Working Paper No. 2017-A1. doi:10.2139/ssrn.3048994

Zhang, D., Smucker, M.D. and Clarke, C.L.A. (2024) "Overview of the TREC 2024 Lateral Reading Track", Proceedings of the 33rd Text REtrieval Conference. https://trec.nist.gov/pubs/trec33/papers/Overview_lateral.pdf

Zhang, D., Smucker, M.D. and Clarke, C.L.A. (2025) "Overview of the TREC 2025 DRAGUN Track: Detection, Retrieval, and Augmented Generation for Understanding News", The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025), NIST Special Publication. https://trec.nist.gov/pubs/trec34/papers/Overview_dragun.pdf