From Questions to Trust Reports: A LLM-IR Framework for the TREC 2025 DRAGUN Track

Ignacy Alwasiak
ignacy.alwasiak@student.uj.edu.pl
Jagiellonian University
Krakow, Poland

Kene Nnolim
knnolim@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Jaclyn Thi
jthi@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Samy Ateia
samy.ateia@ur.de
University of Regensburg
Regensburg, Germany

Markus Bink
markus.bink@ur.de
University of Regensburg
Regensburg, Germany

Gregor Donabauer
gregor.donabauer@ur.de
University of Regensburg
Regensburg, Germany

David Elsweiler
david.elsweiler@ur.de
University of Regensburg
Regensburg, Germany

Udo Kruschwitz
udo.kruschwitz@ur.de
University of Regensburg
Regensburg, Germany

Abstract

The DRAGUN Track at TREC 2025 targets the growing need for effective support tools that help users evaluate the trustworthiness of online news. We describe the UR_Trecking system submitted for both Task 1 (critical question generation) and Task 2 (retrieval-augmented trustworthiness reporting). Our approach combines LLM-based question generation with semantic filtering, diversity enforcement using clustering, and several query expansion strategies (including reasoning-based Chain-of-Thought expansion) to retrieve relevant evidence from the MS MARCO V2.1 segmented corpus. Retrieved documents are re-ranked using a monoT5 model and filtered using an LLM relevance judge together with a domain-level trustworthiness dataset. For Task 2, selected evidence is synthesized by an LLM into concise trustworthiness reports with citations. Results from the official evaluation indicate that Chain-of-Thought query expansion and re-ranking substantially improve both relevance and domain trust compared to baseline retrieval, while question-generation performance shows moderate quality with room for improvement. We conclude by outlining key challenges encountered and suggesting directions for enhancing robustness and trustworthiness assessment in future iterations of the system.

Keywords

Understanding News, Retrieval Augmented Generation, TREC

1 Introduction

The proliferation of misinformation and disinformation across large-scale social networks and news outlets is a complex socio-technical phenomenon [4, 14, 17], with significant implications for how individuals form and adjust their opinions [8]. Existing interventions aim to mitigate these risks, for example, through algorithmic filtering [3] of inappropriate content, external fact-checking services [7], or community-driven initiatives such as community notes [6, 15]. However, such measures are limited to specific platforms and contexts. Once misleading claims circulate beyond these controlled environments, users are left without institutional or algorithmic safeguards and must independently assess the credibility of the information they encounter.

Although many people believe they are good at discerning truth from fiction, studies show that users are often overconfident in their actual evaluation capabilities and thus ill-equipped [11]. Similarly, search behavior can reinforce user-belief in misleading news [1, 5]. This highlights the urgent need to support users in their information search and evaluation practices.

The **DRAGUN** (Detection, Retrieval, and Augmented Generation for Understanding News) Track at TREC 2025 [16] addresses this challenge by proposing two related tasks: Task 1 focuses on generating critical and investigative questions that prompt readers to consider aspects such as source bias, underlying motivations, and diversity of viewpoints when evaluating online content. Task 2 complements this by generating detailed, retrieval-augmented (RAG-based) reports that help users thoughtfully assess the credibility of specific webpages.

Together, these question prompts and analytical reports aim to strengthen users' capacity to discern the trustworthiness of online information and make more informed judgments about the veracity of central claims in news articles.

As part of this paper, we present our participation in the two tasks by describing the resources and methodology we used along with the results we achieved in the official evaluation phase.

To support the reproducibility of our work, we publicly make available all implementations related to our submission on Github: https://github.com/doGregor/UR_trecking_2025.

2 Methods

This section presents our end-to-end methodology (for an overview see Figure 1). We first describe the data, indexing setup, and language models used. We then outline our approach to generating and filtering critical questions, followed by retrieval with query expansion, re-ranking, and credibility-based filtering. Finally, we explain how the resulting evidence is synthesized into question-level answers and a final trustworthiness report.

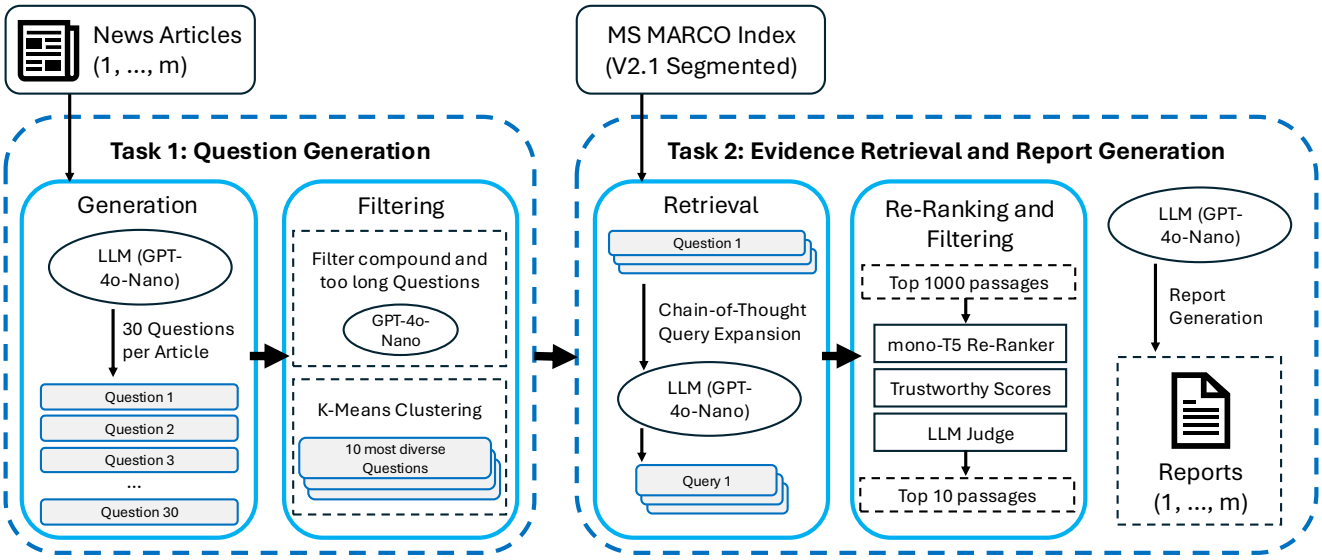

Figure 1: Flowchart of our pipeline for Tasks 1 and 2.

2.1 Data and Resources

We indexed *MS MARCO V2.1 (Segmented)*[12] in OpenSearch using the built-in default English analyzer. We use OpenAI's GPT4o-Nano for LLM-based components in our pipeline. This model was chosen primarily for its tradeoff between costs and efficiency.

2.2 Question Generation and Filtering

To generate questions, we used the aforementioned GPT model with a detailed prompt specifying the aspects we wanted the LLM to focus on. These included logical principles that the LLM should follow when generating questions and question formatting stipulations. For all details of this processing pipeline, including prompts and hyper-parameter settings, please consult the project repository[1].

The following is an excerpt of the prompt provided to the model:

> **Follow these key principles for question generation:**
> 1) Investigate the Source:
> - Generate questions about the publisher's (or key information sources such as organizations, experts, reporters, etc.) background, reputation, and potential biases.
> - Ask about their credentials, expertise, and past work.
> - Inquire about ownership, funding sources, and editorial policies.
>
> ...

Once we had a set of questions generated by GPT, we applied semantic filtering to rid it of invalid questions; these included compound questions and questions that were too long.

Finally, in order to ensure that our set of questions was diverse and covered a broad range of trustworthiness concerns, we applied unsupervised K-means Clustering on embeddings created for each question in our dataset via a sentence transformer. We then chose the question closest to each centroid (in embedding space) found after termination of the algorithm on each set of questions, divided by article, as the questions to keep for each provided article.

In order to evaluate the quality of the questions, we have calculated the TF-IDF score, Jaccard score, and cosine similarity between the question and the article in question. Additionally, we used a LLM to give scores to each question and check how well each question aligns with each CRAAP test[2] component, indicating the reliability and credibility of sources based on currency, relevant, authority, accuracy and purpose. After manual inspection and calculations, we have found that there has been little correlation between the scores given by LLMs.

2.3 Retrieval

After generating critical questions for each target article, we issued these questions as queries to retrieve supporting evidence from the MS MARCO V2.1 segmented corpus. The purpose of this step was twofold: (1) to obtain candidate documents that could be used to answer each question, and (2) to collect citation-worthy evidence for use in the final trustworthiness report.

To improve retrieval quality beyond the baseline (using the original question as the query), we experimented with several query expansion strategies. Specifically, we evaluated the following methods:

- **Baseline (No Expansion):** The original question was issued directly as a search query using BM25.
- **Boolean Expansion:** Additional keywords and phrases were generated and combined with the original query using Boolean operators (e.g., AND/OR) to broaden lexical coverage.

[1] https://github.com/doGregor/UR_trecking_2025

- **Chain-of-Thought (CoT) Expansion:** Following the approach described by Jagerman et al [9], we prompted a large language model to reason step-by-step about the information needed to answer a question and to generate expanded query terms accordingly.
- **Structured Expansion:** Instead of producing a single expanded query string, the language model generated a full OpenSearch Query DSL object. This allowed for more explicit control over field weighting and query composition.

To evaluate retrieval quality, we developed an LLM-based relevance judge to determine whether a retrieved document was judged as relevant to the given question. For each method, we examined the top 10 retrieved documents from the set of $k = 1000$ candidates. A relevance score was computed as the proportion of these top documents that were labeled as relevant by the LLM.

2.4 Re-ranking and Filtering

We run multiple reranking and filtering steps on the list of retrieved documents.

First, we use the mono T5 reranker[2] [13] as fine-tuned on the MS MARCO V2.1 passage dataset to rerank the first 1000 retrieved documents.

We used the LLM-based relevance evaluator described in Section 2.3 to determine whether each retrieved document meaningfully contributed to answering its corresponding question.

To assess the credibility of document sources, we incorporated the domain quality dataset introduced by Lin et al [10]. It combines multiple standard datasets, such as NewsGuard, into a single comprehensive dataset. This dataset assigns a continuous trustworthiness score to news and web domains based on large-scale quality assessments.

For each retrieved document, we extracted its domain from the URL and looked up its corresponding trustworthiness score in the Lin et al. dataset. If a domain was not present in the dataset, it was assigned a default trust score of 0.0. These domain scores were used as an additional signal for filtering and analysis.

After retrieval, query expansion, re-ranking, and relevance evaluation, we applied two final filtering strategies:

- **Top-10 Relevant:** The top 10 documents labeled as relevant by the LLM.
- **Top-3 Relevant and Trustworthy:** The top 3 documents that were both labeled as relevant and had a domain trust score of at least 0.7.

Filtering was applied over the top 100 documents produced after re-ranking. If fewer than the required number of documents satisfied the relevance or trustworthiness constraints, all qualifying relevant documents within the top 100 were retained. This final subset of documents was used for evidence synthesis and citation in the trustworthiness report.

This combined relevance-and-trust filtering process enabled us to focus on documents that were both semantically useful and originated from higher-quality information sources.

[2]https://huggingface.co/castorini/monot5-base-msmarco

2.5 Answer and Report Generation

The <**Question, List of Related Articles**> tuples are analyzed by an LLM to create a sentence with explanation for each question that includes citations. After answering all of the questions an LLM shortens it into a report of reasonable length.

3 Evaluation

For evaluation, NIST assessors independently researched each news article's trustworthiness and created question-and-answer rubrics capturing what a good trustworthiness report should address. One primary assessor consolidated the three assessors' rubrics into a single rubric, which the DRAGUN organizers reviewed and for some cases lightly edited.

All assessor rubric question received an **importance level**:

- Have to Know (4 points)
- Good to Know (2 points)
- or Nice to Know (1 point)

To score the generated questions, two LLMs (Qwen3-Embedding-8B and Qwen3-Reranker-8B) selected the system question most similar to each rubric question. Human assessors then judged the selected pair(s) with one of four **similarity labels**:

- Very Similar (1 point)
- Similar (0.5 points)
- Different (0 points)
- or Very Different (0 points)

For each rubric question, the system earned the rubric question's importance weight multiplied by the highest similarity score available from its matched question(s). The final article-level score was the average of these weighted scores across all rubric questions.

4 Results

This section presents the results for Task 1 (question generation) and Task 2 (retrieval and report generation). We summarize the submitted run and analyze question quality, retrieval effectiveness under different query-expansion strategies, and the impact of re-ranking and trust filtering on final reports.

4.1 Runs Submitted

We submitted a single run for each task (Task 1 and Task 2). The submitted runs included all major components of our pipeline (LLM-based question generation, semantic filtering, clustering-based question selection, retrieval with multiple query-expansion strategies, re-ranking, and trust filtering). However, the iterative correction loop described in Section 2.2 was **not** included in the final run due to time constraints. All remaining components were executed as described.

4.2 Question Generation (Task 1)

Across topics, our question-generation module produced between 10 and 20 candidate questions per article, depending on how many were removed through semantic and compound-question filtering. The official TREC evaluation assigned *moderate* overall qgen_scores, with substantial variation across topics (Table 1).

We consistently observed the following patterns:

- **Very low contradictory scores**, indicating that our generated questions rarely conflicted with the assessor rubrics.
- **Low supportive scores**, suggesting that many questions did not strongly align with what human assessors considered most helpful for trustworthiness assessment.
- **High variance across topics**, with some articles scoring near zero and others showing noticeably stronger alignment (e.g., topics 35 and 41).

Overall, the results indicate that our pipeline effectively avoided generating harmful or contradictory questions, but struggled to consistently produce deeply supportive or rubric-aligned ones. This is likely due to the absence of the planned correction loop and instability in LLM-based scoring during development.

4.3 Retrieval and Report Generation (Task 2)

To evaluate retrieval quality, we compared all query-expansion approaches—baseline, Boolean, Chain-of-Thought (CoT), and Structured—before and after re-ranking. The results, visualized in Figures 1 and 2, show clear and consistent trends:

- **CoT expansion achieved the highest relevance scores** across the first 20 TREC example questions and exhibited the **lowest variance** in performance.
- Boolean and Structured expansions displayed **higher variability**, suggesting greater sensitivity to prompt formulation and query composition.
- **Re-ranking consistently improved both relevance and domain trustworthiness** for all expansion strategies (Figures 3 and 4), demonstrating that the monoT5 reranker promotes not only semantically appropriate documents but also higher-quality domains.

After applying LLM-based relevance labeling and domain-trust filtering (threshold ≥ 0.7), we generated the final RAG-based trustworthiness reports. Supportive scores for these reports were generally low (Table 1), reflecting that the retrieved evidence often addressed article claims only partially. Nonetheless, the stable improvements in both relevance and trust indicate that our retrieval pipeline meaningfully enhanced the quality of sources provided to the report-generation component.

In summary, the results show that **reasoning-based query expansion combined with re-ranking** delivered the most reliable retrieval performance, while question and report generation remained limited by components not fully integrated into the submitted pipeline.

5 Discussion

Our results show that Chain-of-Thought based query expansion combined with monoT5 re-ranking yielded the best trade-off between relevance and domain trust scores for Task 2. This supports the use of reasoning-oriented expansion as a way to expose missing facets in news-related information needs. In contrast, question generation and report writing lagged behind, with low supportive scores, high topic-level variance, and inconsistencies in LLM-based scoring. These issues were amplified by our reliance on LLM components for internal evaluation and relevance judgments, which introduces calibration and alignment uncertainty.

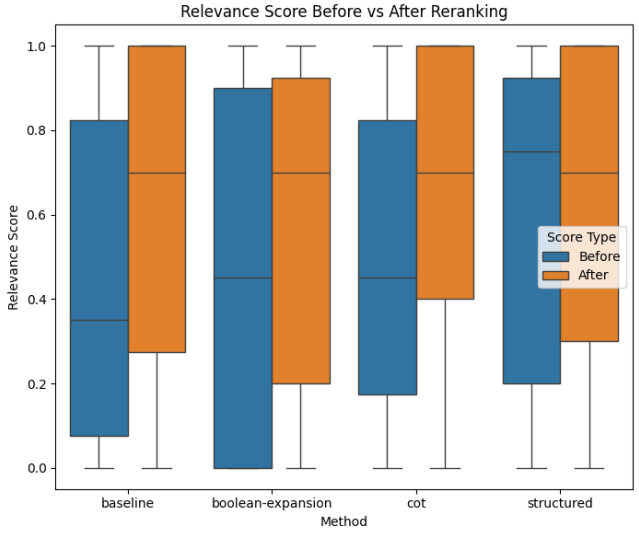

Figure 2: Relevance scores for the first 20 example TREC questions before and after re-ranking across all four query expansion methods.

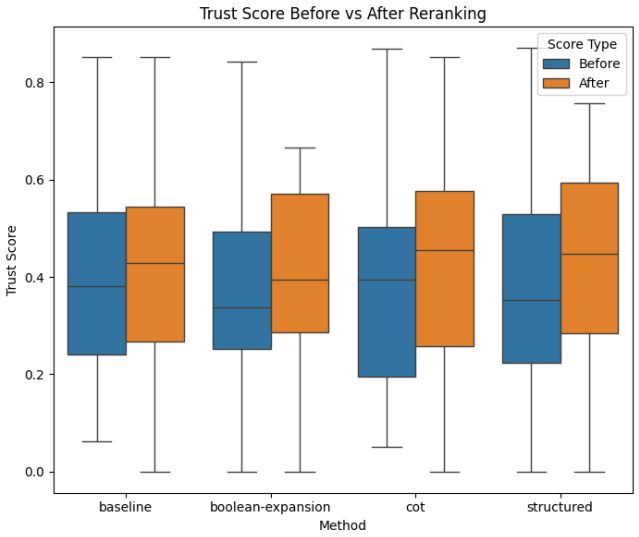

Figure 3: Domain trustworthiness scores for the first 20 example TREC questions before and after re-ranking across all four query expansion methods.

A key limitation of our study is that we submitted only a single run and were unable to integrate the planned iterative correction loop or more advanced question-quality filtering into the final pipeline. As a result, parts of the system remained prototype-level and not tuned against the official evaluation setup. Taken together, the findings suggest that future work should focus on stabilizing question generation, improving LLM-based filtering and scoring, and strengthening the coupling between retrieval, trust signals, and downstream report structure.

topic_id	qgen_score original	qgen_score updated	contradictory_score	supportive_score
msmarco_v2.1_doc_04_420132660	0.259	0.250	0.071	0.000
msmarco_v2.1_doc_06_1440134319	0.350	0.333	0.000	0.000
msmarco_v2.1_doc_08_300872161	0.158	0.100	0.000	0.075
msmarco_v2.1_doc_15_116067546	0.200	0.192	0.051	0.038
msmarco_v2.1_doc_21_861891150	0.226	0.219	0.031	0.312
msmarco_v2.1_doc_22_1648697797	0.250	0.240	0.000	0.053
msmarco_v2.1_doc_25_481628070	0.105	0.100	0.000	0.125
msmarco_v2.1_doc_25_501708725	0.000	0.000	0.000	0.000
msmarco_v2.1_doc_25_502424913	0.174	0.167	0.000	0.000
msmarco_v2.1_doc_34_2529216	0.132	0.129	0.000	0.162
msmarco_v2.1_doc_34_7751734	0.095	0.091	0.000	0.076
msmarco_v2.1_doc_35_1300032609	0.450	0.429	0.000	0.114
msmarco_v2.1_doc_35_441326441	0.133	0.125	0.000	0.000
msmarco_v2.1_doc_38_227081897	0.353	0.111	0.000	0.000
msmarco_v2.1_doc_39_1014551192	0.059	0.000	0.000	0.028
msmarco_v2.1_doc_39_1165221603	0.077	0.074	0.049	0.019
msmarco_v2.1_doc_41_1960853470	0.433	0.125	0.000	0.075
msmarco_v2.1_doc_42_654618974	0.318	0.292	0.000	0.028
msmarco_v2.1_doc_47_1430382251	0.286	0.273	0.000	0.114
msmarco_v2.1_doc_48_1289995263	0.107	0.103	0.000	0.000
msmarco_v2.1_doc_48_273997000	0.000	0.000	0.000	0.057
msmarco_v2.1_doc_48_515083157	0.087	0.083	0.083	0.219
msmarco_v2.1_doc_48_515287844	0.077	0.000	0.000	0.190
msmarco_v2.1_doc_48_730773621	0.471	0.444	0.000	0.111
msmarco_v2.1_doc_52_1666095905	0.088	0.028	0.102	0.296
msmarco_v2.1_doc_54_1547893878	0.000	0.000	0.000	0.000
msmarco_v2.1_doc_54_304836636	0.114	0.109	0.000	0.009
msmarco_v2.1_doc_55_248024230	0.038	0.037	0.000	0.000
msmarco_v2.1_doc_57_933363597	0.308	0.286	0.000	0.214
msmarco_v2.1_doc_58_748655897	0.000	0.000	0.000	0.029

Table 1: Results of our submitted run for question generation (*qgen_score original* represents the scores of the original evaluation while *qgen_score updated* are updated evaluation results the DRAGUN organizers sent after slightly changing the evaluation setup for task 1) as well as contradictory score (the lower, the better) and supportive score (the higher, the better) for generated report.

6 Conclusion

In this paper, we presented UR_Trecking, a retrieval-augmented LLM-based pipeline for critical question generation and trustworthiness reporting in the DRAGUN Track. Our main contribution is an empirical analysis of Chain-of-Thought query expansion plus neural re-ranking augmented with domain-level trust signals, together with an end-to-end implementation for news trustworthiness support on MS MARCO V2.1 segmented. The evaluation indicates gains in retrieval relevance and domain trustworthiness, but only moderate alignment of generated questions and reports with assessor rubrics. Next steps are to complete and stabilize the pipeline by integrating iterative correction, making LLM-based scoring more robust, and explicitly optimizing for rubric coverage. In the longer term, user studies will be needed to test whether these technical improvements translate into better trust judgments in real-world news consumption.

References

[1] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A Tucker. 2024. Online searches to evaluate misinformation can increase its perceived veracity. *Nature* 625, 7995 (2024), 548–556.

[2] Sarah Blakeslee. 2004. The CRAAP test. *Loex Quarterly* 31, 3 (2004), 4.

[3] Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen, and Dosam Hwang. 2021. Trends in combating fake news on social media–a survey. *Journal of Information and Telecommunication* 5, 2 (2021), 247–266.

[4] Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization* 100, 9 (2022), 544.

[5] David Elsweiler, Samy Ateia, Markus Bink, Gregor Donabauer, Marcos Fernández Pichel, Alexander Frummet, Udo Kruschwitz, David E. Losada, Bernd Ludwig, Selina Meyer, and Noel Pascual Presa. 2025. Query Smarter, Trust Better? Exploring Search Behaviours for Verifying News Accuracy. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 515–526. doi:10.1145/3726302.3730067

[6] William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety* 1, 1 (2021).

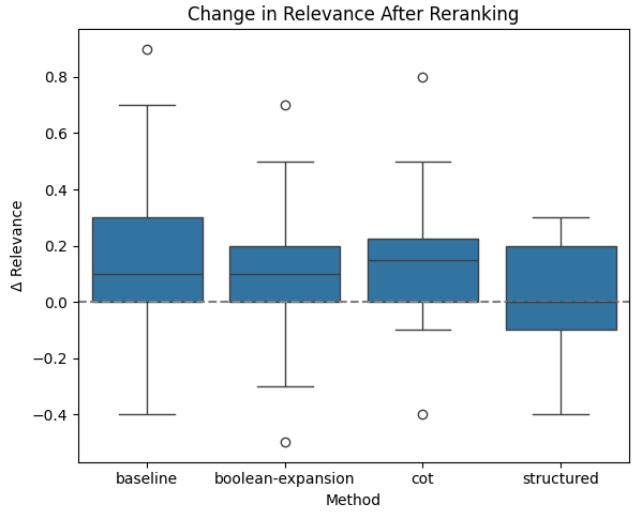

Figure 4: Change in relevance scores after re-ranking for all four query expansion methods.

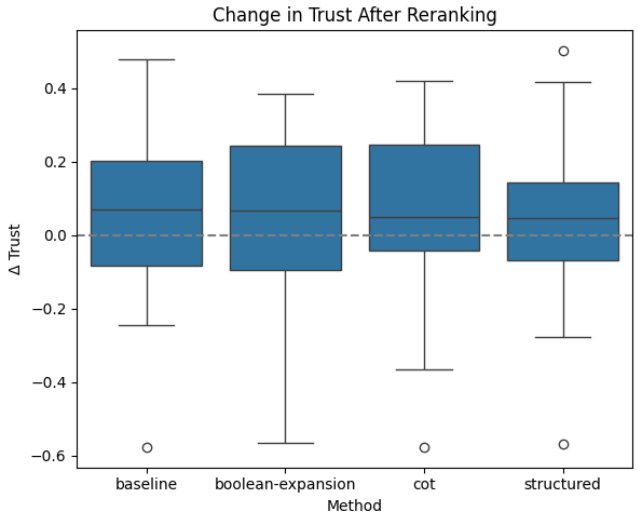

Figure 5: Change in domain trustworthiness scores after re-ranking for all four query expansion methods.

[7] Michael Hameleers and Toni GLA Van der Meer. 2020. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication research* 47, 2 (2020), 227–250.

[8] Effiati Juliana Hasibuan, Alifiansyah Putra, Rahmana Deto, and Annisagita Sungga Dirgantari. 2024. The Role of Social Media Algorithms in Shaping Public Opinion During Political Campaigns. *International Journal of Social and Human (IJSH)* 1, 2 (2024), 165–172.

[9] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query Expansion by Prompting Large Language Models. *arXiv preprint* abs/2305.03653 (2023). https://arxiv.org/abs/2305.03653

[10] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G. Rand, and Gordon Pennycook. 2023. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* 2, 9 (Sept. 2023), pgad286. doi:10.1093/pnasnexus/pgad286

[11] Khalid Mahmood. 2016. Do people overestimate their information literacy skills? A systematic review of empirical evidence on the Dunning-Kruger effect. *Communications in Information Literacy* 10, 2 (2016), 3.

[12] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human-Generated Machine Reading Comprehension Dataset. (2016).

[13] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. doi:10.18653/v1/2020.findings-emnlp.63

[14] Stephen Prochaska, Kayla Duskin, Zarine Kharazian, Carly Minow, Stephanie Blucker, Sylvie Venuto, Jevin D West, and Kate Starbird. 2023. Mobilizing manufactured reality: How participatory disinformation shaped deep stories to catalyze action during the 2020 US presidential election. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–39.

[15] Nicolas Pröllochs. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 794–805.

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

[17] Yanmengqian Zhou and Lijiang Shen. 2022. Confirmation bias and the persistence of misinformation on climate change. *Communication Research* 49, 4 (2022), 500–523.