Overview of the TREC 2024 Lateral Reading Track

Dake Zhang, Mark D. Smucker, Charles L. A. Clarke

University of Waterloo, Canada

Abstract

The current web landscape, characterized by abundant information and widespread misinformation, highlights the pressing need for people to evaluate the trustworthiness of online content effectively. However, this remains a daunting challenge for many internet users. The TREC 2024 Lateral Reading Track seeks to address this issue by supporting the use of *lateral reading*, a proven strategy used by professional fact-checkers, to help users evaluate news articles more effectively and efficiently. In its first year, the track had two tasks: (1) generating questions that readers should consider when assessing the trustworthiness of the given news articles, and (2) retrieving documents to help answer these questions. This paper presents an overview of the track, including its objectives, methodologies, resources, and evaluation results. Our evaluation of the submitted runs shows the significant challenges these tasks pose to existing approaches, including state-of-the-art large language models. Further details on this track can be found on its website: `https://trec-dragun.github.io/`.

1 Introduction

Evaluating the trustworthiness of information is a task that remains challenging for many users [5]. A study by Wineburg and McGrew [8] revealed that professional fact-checkers employ a systematic approach named *Lateral Reading*, which involves examining source credibility, supporting evidence, and cross-referencing other sources through web searches beyond the web page to be assessed. Their use of lateral reading enabled them to achieve higher accuracy and faster trustworthiness evaluation compared with university students and historians, who spent most of their time looking for superficial credibility indicators within the given web page, i.e., *vertical reading*.

The TREC 2024 Lateral Reading Track is designed for researchers interested in helping people assess the trustworthiness of information. Rather than explicitly defining truth or misinformation, this track aims to help users make informed decisions by providing comprehensive background and contextual information.

In its first year, the track had two tasks: 1) Question Generation and 2) Document Retrieval, which we believe are the key components for any system that will support lateral reading. The first task required participants to generate questions that readers should consider when evaluating the trustworthiness of the given news articles. The second task focused on retrieving relevant documents from `ClueWeb22-B-English` to answer those questions.

In the following sections, we present an overview of the task definitions, describe the resources, explain our assessment methodology for participant submissions, and report the evaluation results. Our analysis of submitted runs shows the difficulties that these tasks present to current approaches, including current Large Language Models (LLMs), and as such, shows there is still considerable room for improvement in our ability to support people's trustworthiness assessment of information.

2 Methods and Materials

The two tasks in the first year had separate submission deadlines. In Task 1: Question Generation, participants needed to generate questions that readers of an online news article should ask when evaluating its trustworthiness. After the deadline of Task 1, we pooled some questions submitted by participants as search queries for Task 2: Document Retrieval. Participants in Task 2 needed to retrieve documents from the specified document collection `ClueWeb22-B-English` that would contain answers to those questions, in the context of the given news articles. Task 2 is similar to traditional ad-hoc retrieval tasks.

2.1 Data

This track used the Category B dataset from the English subset of `ClueWeb22` [7] as its document collection, which has approximately 87 million frequently accessed web documents crawled in February 2022. We used the plain-text version of those documents.

To construct the news article set, we first identified 50 widely discussed news events from 2021 and early 2022. For each event, we retrieved one representative news article from the document collection, while ensuring diverse media coverage across 50 events. The selected articles came from 43 distinct media outlets, with their political leanings distributed as follows: 26 from left-leaning sources, 16 from right-leaning sources, 4 from neutral sources, 2 from pro-science sources, and 2 from sources without established bias ratings, according to Media Bias/Fact Check[1]. Each selected news article served as a *topic* for this track. We made the 50 articles available via both a list of document identifiers and as part of our `ClueWeb22` distribution `TREC-LR-2024`[2], which CMU distributed.

2.2 Tasks

The scenario assumes a general public reader browsing online news. The specific requirements for each task are outlined below.

2.2.1 Task 1: Question Generation

For each of the 50 topics (i.e., target news articles), participants needed to produce **10** questions that the reader should ask to evaluate its trustworthiness, ranked from the most important to the least important. Participants were told that each question should meet the following requirements:

- Should be self-contained and explain the full context, i.e., one can understand this question without reference to the article.
- Should be at most 120 characters long.
- Should be reasonably expected to be answered by a single web page.
- Compound questions should be avoided, e.g. who is X and when did Y happen? In general, each question should focus on a single topic.

Submissions could be either *manual* (involving human intervention to generate questions, e.g., hiring people to produce questions or manually selecting questions from a candidate list of questions produced by algorithms) or *automatic* (automatic systems that produce questions without the need

[1] `https://mediabiasfactcheck.com/`

[2] `https://lemurproject.org/clueweb22/obtain.php`

of human input beyond the construction of the systems). As usual with TREC, teams submitting automatic runs were to make a good-faith effort not to read or study the 50 articles.

To illustrate what kind of questions were expected, we provided 10 example questions for a news article selected outside of the document collection. On February 21, 2023, the New York Times published an opinion article by Bret Stephens entitled "*The Mask Mandates Did Nothing. Will Any Lessons Be Learned*". In the article, Stephens made an argument that mask mandates during the COVID-19 pandemic did not work. Given the importance of this issue, the reader would be advised to examine the trustworthiness of the information. As suggested by lateral reading, questions should be asked about information sources, the evidence presented, and what others say about the issue. Below are the 10 questions we manually created to evaluate the trustworthiness of this article, based on its plain-text content. Note that we noticed later that some of the example questions are longer than 120 characters in order to be self-contained.

- Are reviews by Cochrane, a British non-profit, a reliable source of health care data?
- Did Cochrane, a British non-profit, publish a study in 2023 indicating that mask mandates are not effective for reducing the spread of respiratory illnesses — including COVID-19?
- Could Tom Jefferson, the Oxford epidemiologist, be considered an expert on mask mandates and the spread of respiratory illnesses?
- Could Rochelle Walensky, director of the Centers for Disease Control and Prevention, be considered an expert on mask mandates and the spread of respiratory illnesses?
- What evidence is there that wearing masks can protect against respiratory illnesses — including COVID-19?
- Are N-95 masks better than lower-quality surgical or cloth masks at protecting against respiratory illnesses — including COVID-19?
- What is the guidance from the Center for Disease Control and Prevention on mask mandates in schools?
- What are the political leanings of Bret Stephens, the New York Times opinion columnist?
- What are the political leanings of the journalist Maryanne Demasi?
- Does China provide a highly successful model for pandemic response?

In working to answer these questions, the reader would likely learn that Bret Stephens is a conservative [1], that Tom Jefferson had previously published articles using other studies as evidence against masks, which received criticism from other scientists [3], that Maryanne Demasi is a journalist who has faced criticism for reports that go against scientific consensus, e.g. Wi-Fi is dangerous [2], and that the Cochrane study was misinterpreted as it was inconclusive about the question of if interventions to encourage mask wearing worked or not [4]. With this information in hand, the reader should be better able to judge the trustworthiness of this news article.

For Task 1, we received 19 runs from 4 groups, including 2 baseline runs that we created.

2.2.2 Task 2: Document Retrieval

Our original plan was to have assessors manually write questions for the 50 news articles and use those questions as search queries for Task 2. Unfortunately, assessors did not start working on this track until late August. So in late July, we decided to pool some questions from the Task 1 submissions as search queries for this document retrieval task.

For each target news article, we pooled 3 questions from each of the 4 groups that participated in Task 1, resulting in a total of 12 questions. The selection process followed a randomized round-robin approach. First, we randomized the order of the groups, then proceeded by iteratively selecting one question from the top prioritized run in that group (based on the priority specified in the submission form). During each turn, we used `GPT-4-128k` (2024-04-09 version) to assess whether the current question met the Task 1 question requirements and was at least somewhat helpful for readers. If a question did not satisfy these criteria, we moved to the next question in the same run, and then the second prioritized run if the top one was exhausted. We also used `GPT-4-128k` to check for duplication to avoid selecting questions that were too similar to those already chosen.

In Task 2, for each of the 12 questions, participants needed to produce a ranking of **10** documents from the document collection (`ClueWeb22-B-English`) such that the document was predicted to be useful for answering the question in the context of the corresponding news article. This task is similar to a traditional ad-hoc retrieval task, except for the article context. Participants were free to use the target news articles in addition to the pooled questions during their retrieval processes.

Runs could be either full-rank or rerank. For the rerank option, we prepared a BM25-RM3 baseline run (`Organizers-Baseline-BM25RM3`) that retrieved the top 100 results using BM25 (`k1=0.9, b=0.4`) with RM3 (`fb_terms=10, fb_docs=10, original_query_weight=0.5`) as implemented in Pyserini[3]. We also made the BM25-RM3 run with the content of retrieved documents available as part of our `ClueWeb22` distribution `TREC-LR-2024`[4], which CMU distributed. Participants could choose to rerank those results without the hassle of indexing the full collection. Similar to Task 1, runs could be either automatic or manual.

We received 10 runs from 4 groups for this task, including 6 full-rank runs, 2 rerank runs, and 2 baseline (full-rank) runs from us.

2.3 Organizer Baselines

As organizers, we constructed two automatic runs for the question generation task, by prompting `GPT-4`. The prompts were based on the instructions we provided to assessors and aimed to evaluate the performance of LLMs in generating questions without extensive prompt engineering. For the document retrieval task, as noted earlier, we prepared a baseline run using BM25 with RM3 for participants to rerank. Additionally, we created an enhanced run by incorporating an LLM-based query expansion component and an LLM-based judge component. This adjustment was designed to make the retrieved documents more relevant to those questions in the context of news articles.

2.3.1 Automatic Question Generation

Our two automatic runs (`Organizers-Baseline-1` and `Organizers-Baseline-2`) for the question generation task were created using `GPT-4-128k` (2024-04-09 version). The system prompt came from the instructions[5] we provided to NIST assessors, specifically from Parts I to IV. The only difference between the two runs lies in the inclusion of examples in the prompt. For `Organizers-Baseline-2`, the prompt included the example news article and sample questions, as described in Section 2.2.1. These examples correspond to Part V of the assessing instructions.

In both runs, the news article was inserted into the user input, followed by a reiterated task instruction to generate 10 questions aimed at helping readers evaluate the trustworthiness of the

[3] `https://github.com/castorini/pyserini`

[4] `https://lemurproject.org/clueweb22/obtain.php`

[5] `https://trec-lateral-reading.github.io/assessing_instructions.pdf`

article. The generated questions were then double-checked by code to ensure none exceeded 120 characters in length.

For these runs, the objective was to show how an LLM performs when given instructions identical to those provided to human assessors, so we did not explore advanced prompt engineering techniques. Additionally, the order of the questions was the exact output sequence generated by the model without further adjustments.

2.3.2 Document Retrieval

In addition to the baseline run, `Organizers-Baseline-BM25RM3`, introduced in Section 2.2.2, which we provided for participants to rerank, we also created a more advanced retrieval run (`Organizers-LLM-Assessor`), using `GPT-4o` (2024-08-06 version) for query expansion, pooling retrieved documents through BM25 across expanded queries, and assessing their usefulness using `Llama-3.1-8B-Instruct`.

For the query expansion, we used `GPT-4o` with a system prompt instructing it to act as an intelligent search engine assistant. The model was tasked to generate 10 distinct and effective queries tailored for term-based search algorithms (e.g., BM25) to retrieve relevant documents, based on the initial question. Both the initial question and the associated news article were provided in the user input to guide the query expansion.

For each of the 10 expanded queries, we used BM25 (`k1=1.2, b=0.75`) to retrieve the top 30 documents. We then assessed the usefulness of each unique document across 10 searches using `Llama-3.1-8B-Instruct`, selected for its cost-efficiency and computational speed compared to larger variants, albeit with a potential trade-off in performance. In its system prompt, the model was directed to act as a professional text analyst, responsible for assessing the usefulness of each document in addressing the initial question, with three levels: Very Useful [2], Useful [1], and Not Useful [0]. These definitions aligned with the assessing instructions for human assessors detailed in Section 2.4.3.

To obtain the final ranked list, we assigned each document a composite score, by summing the usefulness score (multiplied by 1000) and the aggregated BM25 scores across all 10 expanded queries. This scoring scheme prioritized the usefulness assessment from `Llama-3.1-8B-Instruct` while also considering document relevance indicated by high BM25 rankings across multiple expanded queries. The rationale behind this approach was that documents assessed as useful by the model and consistently retrieved across diverse queries were more likely to be relevant.

2.4 Run Assessment

Each topic was assigned one primary assessor, who was responsible for writing questions, assessing participant questions, and assessing retrieved documents. Our assessing instructions are available at: `https://trec-lateral-reading.github.io/assessing_instructions.pdf`

2.4.1 Assessor Questions

To help assessors understand Task 1, for each topic (article), the assigned assessor was asked to scrutinize the article in plain text and produce 10 questions that the reader should ask to evaluate its trustworthiness, ranked by their importance to the evaluation from the most important to the least important. The requirements for these questions were the same as those mentioned in Section 2.2.1, except that no strict length requirement was imposed to reduce the cognitive load on assessors. If the article contains reader comments at the bottom of the page (there should be few if any of these), assessors were to ignore the reader comments and read only the main article

to formulate their questions. We also included some background knowledge of lateral reading in our instructions to assessors.

Besides the primary assessor, each topic was assigned two additional (secondary) assessors to have more lists of questions. However, some assessors ran out of time writing questions for all their allocated secondary topics, and therefore, some articles have fewer than three sets of questions. All these collected assessor questions have not been used in evaluating submissions this year. We hope to utilize these questions in next year's track, and as such, we have not released them at this time.

2.4.2 Question Assessment

After writing their questions, assessors were asked to assess the helpfulness of the questions submitted by participants. They were presented with a ranked list of 10 questions from a run each time. Their assessments were on two aspects: *quality* and *redundancy*.

Their *quality* assessments came from the following grades.

- Flawed [-1]:
  - The question cannot be used as an independent query without reference to the article, i.e., not self-contained (contextualized). Example: *Does the article show any political or financial bias?*
  - OR The question is not related to the article.
  - OR The question has other issues that are not listed above and therefore cannot be comprehended.
- Not Helpful [0]:
  - The question does not help in evaluating the trustworthiness of the article, even if it is somewhat related.
  - The question may be compound or not be expected to be answered by a single web page.
- Okay [1]:
  - The question provides helpful information for understanding topics in the article but is not crucial for readers to judge trustworthiness. In other words, it might be helpful for some readers, but not important for most readers during their trustworthiness evaluation.
  - The question may be compound or not be expected to be answered by a single web page.
- Good [2]:
  - While not critical, an answer to the question will enhance readers' confidence in their judgment of the article's trustworthiness.
  - The question may be compound or not be expected to be answered by a single web page.
- Very Good [3]:
  - The question addresses the core aspects of the article's trustworthiness. An answer to the question will be critical for readers to form reliable judgments about the article's trustworthiness. Its answer can potentially flip readers' trustworthiness perception of the article.
  - The question may be compound or not be expected to be answered by a single web page.
- Excellent [4]:

– The question addresses the core aspects of the article's trustworthiness. An answer to the question will be critical for readers to form reliable judgments about the article's trustworthiness. Its answer can potentially flip readers' trustworthiness perception of the article.
– AND You could expect the question to be answered by a single web page, even if the question is compound. For example, "*Where does Bret Baier work and what is his job?*" could be expected to be answered in a single web page.

Redundancy refers to whether an answer to a question earlier in the ranked list (within one run) largely suffices as an answer to the current question. For example, the questions "*What is the credibility of Sheera Frenkel in reporting on Middle Eastern conflicts?*" and "*Is Sheera Frenkel a reliable journalist with expertise in Middle Eastern conflicts?*" should be considered redundant. However, upon examining the assessors' redundancy labels, we found that some assessors appeared to accidentally assess redundancy across different question lists, rather than within one question list. Due to this inconsistency, we decided to exclude these labels from our evaluation of the runs for Task 1.

2.4.3 Document Assessment

We pooled retrieved documents from all runs for document assessment. Due to the limited budget, assessors could not complete assessing retrieved documents for all 12 questions of each article. Thus, we asked assessors to select 5 questions whose answers they think a general reader would find most helpful for judging the trustworthiness of the article. If they had extra time, they would assess retrieved documents for more than 5 questions per article.

Documents were assessed with regard to their *usefulness* in answering the given question, in the context of the corresponding news article. This is similar to the evaluation of traditional ad-hoc retrieval tasks, with the following grades.

- Very Useful [2]: The document is very useful for answering the question because it directly addresses the question with an explicit and complete answer or an answer can be derived easily from this document alone.
- Useful [1]: The document is useful for answering the question because it contains useful information. Other documents might be needed to derive a reliable answer to the question, i.e., the document itself only contains a partial answer.
- Not Useful [0]: The document is irrelevant to the question or useless for deriving an answer to the question. The document is also considered useless if it is not written in English, contains inappropriate contents (e.g., adult materials), or is unreadable.

In total, 300 questions were assessed. For the evaluation of Task 2 submissions, we removed 42 questions where there were fewer than 3 useful documents or all runs had NDCG@10 scores of 1.

3 Results

In this section, we present the evaluation results of the submitted runs for both tasks. For the detailed implementation of each run, we refer readers to the respective track paper provided by the participating group.

Table 1: Evaluation results for Task 1 (Question Generation) runs, sorted by DCG@10.

Run ID	Average@10	DCG@10	NDCG@10
h2oloo-gpt4o-stepwise-decompose	2.800	13.366	0.735
h2oloo-gpt4o-decompose	2.776	13.189	0.726
h2oloo-gpt4o-stepwise	2.708	12.978	0.714
h2oloo-mistral-large2-decompose	2.584	12.323	0.678
h2oloo-gpt4o	2.548	12.155	0.669
h2oloo-mistral-large2	2.586	12.127	0.667
h2oloo-llama70-stepwise-decompose	2.348	11.339	0.624
h2oloo-llama70-stepwise	2.214	10.823	0.596
h2oloo-llama70-decompose	2.256	10.811	0.595
Organizers-Baseline-2	2.130	10.486	0.577
portiesAutoSystemA	2.266	10.459	0.574
Organizers-Baseline-1	1.882	9.590	0.528
portiesAutoSystemBOmni	1.854	9.213	0.507
portiesAutoSystemB	1.734	8.582	0.472
uwclarke_auto	1.628	8.387	0.461
h2oloo-llama70	1.650	8.071	0.444
portiesAutoSystemAOmni	1.558	7.636	0.420
uwclarke_auto_summarized	1.312	6.939	0.382
portiesAutoSystemCOmni	0.936	4.414	0.243

3.1 Task 1: Question Generation

We used Discounted Cumulative Gain (DCG) as the primary evaluation metric for this task since it captures both the *quality* of the generated questions and their ranking. Specifically, the gain of each question was calculated by dividing its *quality* grade (ranging from -1 to 4, where -1, indicating *flawed* questions, was treated as a gain of 0) by the reduction factor $log_2(p+1)$, where p denotes the rank position of the question in the list. Additionally, we computed the average *quality* grade of the 10 questions (Average@10) and the normalized DCG (NDCG@10). We treated the ideal ranking as consisting entirely of *excellent* (grade 4) questions, and thus, the ideal DCG (IDCG) was constant across topics.

Table 1 presents the evaluation results for the submitted runs, sorted by DCG@10. All the runs were *automatic* through prompting LLMs, as declared by participants in their submission forms. We observed a wide range of performance levels across runs, with a noticeable gap between the best-performing run and the theoretical upper bound. For example, even the best-performing run, `h2oloo-gpt4o-stepwise-decompose`, achieved an NDCG@10 of 0.735, indicating room for improvement in achieving a ranking of all *excellent* questions.

In terms of topic differences, Figure 1 presents the distribution of DCG@10 scores for all submissions across various articles. The results indicate that news articles in this track exhibit differing levels of difficulty for LLMs in generating helpful questions to support readers' trustworthiness evaluation. Specifically, some articles are highly challenging, characterized by consistently low DCG@10 scores across all runs, as shown on the left side of the figure. In contrast, certain articles are comparatively easy, with higher scores appearing on the right side of the figure.

In general, longer articles with extensive citations tend to make it easy to generate good questions, particularly when these questions focus on the credibility of sources, as suggested by lateral

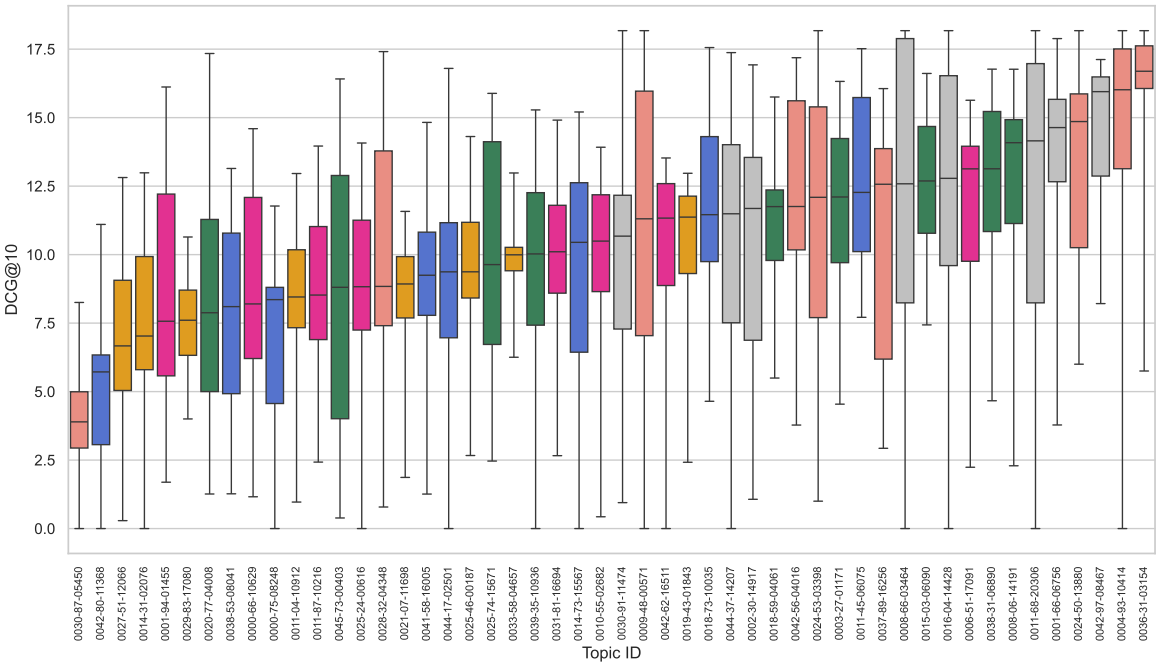

Figure 1: Boxplots of the distribution of per-topic DCG@10 scores, with the minimum, 25th percentile, median, 75th percentile, and maximum values. The topics (article IDs from `ClueWeb22`) are ordered along the x-axis based on their median DCG@10 scores. Different colors distinguish individual assessors.

reading. Meanwhile, some articles from highly credible sources, such as those explaining scientific concepts, pose greater challenges for question generation. In these cases, it is harder to identify aspects that warrant questioning, and readers are often less motivated to examine the trustworthiness of the content. For the next iteration of this track, we plan to include more contentious news articles from sources representing opposite ends of the political or ideological bias spectrum to better explore the question generation task.

3.2 Task 2: Document Retrieval

For the document retrieval task, we evaluated the performance of submitted runs using two widely used metrics in traditional ad-hoc retrieval: NDCG@10 (as the primary metric) and Mean Reciprocal Rank (MRR). The evaluation results are shown in Table 2, with runs ranked by NDCG@10. All runs were declared to be *automatic* in the submission forms. The results indicate that many queries (questions) were relatively easy, as evidenced by the high NDCG@10 and MRR scores achieved across most runs, including the simple baseline run (`Organizers-Baseline-BM25RM3`).

As detailed in Section 2.3.2, we utilized `Llama-3.1-8B-Instruct` to assess the usefulness of documents retrieved by BM25. These LLM-based assessments were compared against those conducted by human assessors. Our analysis revealed a low level of agreement between the model and human judgments, with a Cohen's kappa [6] of just 0.133. Furthermore, compared to human assessments, the model exhibited a tendency to overestimate the usefulness of retrieved documents, leading to a large number of false positives.

As illustrated in Figure 2, a large proportion of the queries have median NDCG@10 scores over 0.5 across all submitted runs. Some relatively easy questions, located on the right side of

Table 2: Evaluation results for Task 2 (Document Retrieval) runs, sorted by NDCG@10.

Run ID	Type	NDCG@10	MRR
h2oloo-bm25-rocchio-monot5-gpt4o	Full-Rank	0.700	0.890
h2oloo-fused-gpt4o-zephyr-llama31_70b	Full-Rank	0.690	0.885
Organizers-LLM-Assessor	Full-Rank	0.651	0.876
h2oloo-bm25-rocchio-monot5-zephyr	Full-Rank	0.629	0.859
h2oloo-bm25-rocchio-monot5	Full-Rank	0.629	0.845
h2oloo-bm25-rocchio-monot5-lit5_xl_v2	Full-Rank	0.610	0.844
h2oloo-bm25-rocchio-monot5-lit5_large_v2	Full-Rank	0.591	0.824
UWClarke_rerank	Rerank	0.547	0.736
h2oloo-bm25-rocchio	Full-Rank	0.521	0.746
Organizers-Baseline-BM25RM3	Baseline	0.511	0.735
TMU_V_BERTSim3	Rerank	0.418	0.637

the figure, achieve a median NDCG@10 score of 1. These questions contain useful keywords that directly lead to relevant documents without requiring reference to the corresponding news articles, as demonstrated by the following examples:

- Is the Indian Express rated green by Newsguard?
- Did Merritt's wastewater treatment plant fail due to flooding?
- Has Wang Yaping been confirmed as the first Chinese woman to conduct a spacewalk?
- Was Ma'Khia Bryant holding a knife during the police shooting?
- What legal authority does Biden have to cancel student loans?
- Did Richard Ayvazyan receive a 17-year sentence for COVID fraud?

In contrast, the most challenging questions, positioned on the left side of the figure, have near-zero NDCG@10 scores. Interestingly, our run, `Organizers-LLM-Assessor`, is the only run that achieved non-zero NDCG@10 scores (over 0.9) on these questions. Despite these questions being seemingly contextualized, term-based retrieval methods struggled with finding relevant documents. Examples of such questions include:

- Does the 100-page report from Moab City exist?
- What is the track record of Live Nation regarding safety at their events?
- What is the official stance of the United States on the new Israeli government?
- Are claims about Biden's involvement in the Great Reset verified?

Taking the first question as an example, incorporating terms such as *Gabby Petito*, *Brian Laundrie*, or *Police* through our query expansion component enabled BM25 to retrieve at least 6 *very useful* documents. This suggests the need for query expansion or rewriting in adapting queries to include the necessary contextual information from the corresponding news articles for this document retrieval task.

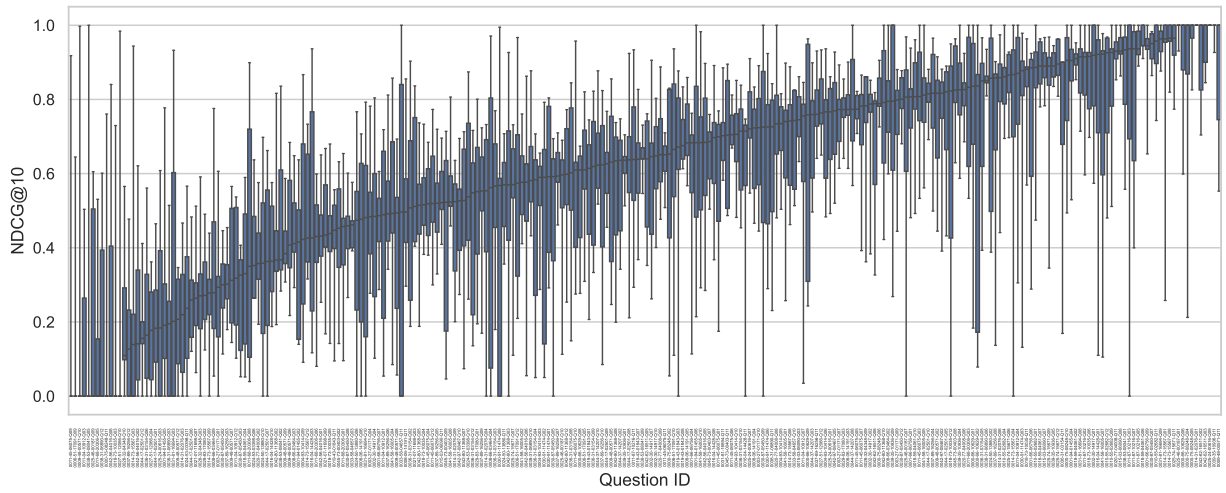

Figure 2: Boxplots of the distribution of per-question NDCG@10 scores, with the minimum, 25th percentile, median, 75th percentile, and maximum values. The question IDs are ordered along the x-axis based on their median DCG@10 scores.

4 Conclusion

In the first year of this track, we designed two foundational tasks that are likely to play an important role in automated systems aimed at supporting the trustworthiness assessment of online news: question generation and document retrieval. Our evaluation of the submitted runs revealed that prompting state-of-the-art LLMs alone is insufficient to generate consistently high-quality questions as judged by human assessors. This finding highlights opportunities for improvement, such as incorporating agentic workflows that utilize web resources to acquire additional context to enhance question generation. Moreover, our experiments showed that query expansion could be beneficial for document retrieval, as it enriches queries with the context of the corresponding news article. Building upon this year's foundation, the track will continue in 2025 under a new name: **DRAGUN** (**D**etection, **R**etrieval, and **A**ugmented **G**eneration for **U**nderstanding **N**ews). The new name aims to encourage broader participation with a stronger emphasis on Retrieval-Augmented Generation (RAG) techniques. The primary task in the next iteration will extend beyond this year's scope, focusing on the generation of attributed reports to support trustworthiness assessment.

Acknowledgments

This work was supported in part by Microsoft and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to express our gratitude to the attendees of TREC 2023 for their insightful discussions on track planning, and to the participants of the 2024 track for their great submissions. Special thanks go to Ian M. Soboroff and Hoa T. Dang of NIST for coordinating the efforts to assess runs, and to the TREC assessors for providing their valuable evaluation. We also thank Diana Brebeanu and Anna Li for their assistance in selecting news articles from `ClueWeb22`.

References

[1] Wikipedia Contributors. Bret Stephens — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Bret_Stephens&oldid=1256439085, 2024. [Online; accessed 13-November-2024].

[2] Wikipedia Contributors. Maryanne Demasi — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Maryanne_Demasi&oldid=1175116938, 2024. [Online; accessed 13-November-2024].

[3] David Gorski. The Cochrane Mask Fiasco: How the Evidence-Based Medicine Paradigm Can Produce Misleading Results. https://sciencebasedmedicine.org/the-cochrane-mask-fiasco-how-the-evidence-based-medicine-paradigm-can-produce-misleading-results/, 2023. [Online; accessed 13-November-2024].

[4] Jessica McDonald. What the Cochrane Review Says About Masks For COVID-19 — and What It Doesn't. https://www.factcheck.org/2023/03/scicheck-what-the-cochrane-review-says-about-masks-for-covid-19-and-what-it-doesnt/, 2023. [Online; accessed 13-November-2024].

[5] Sarah McGrew, Joel Breakstone, Teresa Ortega, Mark Smith, and Sam Wineburg. Can Students Evaluate Online Sources? Learning From Assessments of Civic Online Reasoning. *Theory & Research in Social Education*, 46(2):165–193, 2018. doi: 10.1080/00933104.2017.1416320.

[6] Mary L. McHugh. Interrater Reliability: the Kappa Statistic. *Biochemia Medica*, 22(3):276–282, 2012. doi: 10.11613/BM.2012.031.

[7] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3360–3362, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3536321.

[8] 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. doi: 10.1177/016146811912101102.