Overview of the TREC 2025 DRAGUN Track: Detection, Retrieval, and Augmented Generation for Understanding News

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

University of Waterloo, Canada

Abstract

Many internet users struggle to assess whether online information is trustworthy, a critical skill in today's digital environment where accurate content coexists with false or misleading material. As the successor to the previous TREC 2024 Lateral Reading Track, the TREC 2025 DRAGUN (**D**etection, **R**etrieval, and **A**ugmented **G**eneration for **U**nderstanding **N**ews) Track aims to advance research on supporting readers in assessing the trustworthiness of online news by providing reader-oriented, well-attributed reports. The track had two tasks: (1) Question Generation, which asked participants to propose critical, ranked questions a reader might investigate for a given article; and (2) Report Generation, which asked participants to produce a short (up to 250 words) background report grounded in the MS MARCO V2.1 Segmented Corpus. Using assessor-built rubrics with importance-weighted questions and short answers, we evaluated question coverage and report support/contradiction. We release topics, rubrics, annotations, runs, and evaluation results to support research on developing systems to help people assess the trustworthiness of news.

1 Introduction

This is the second and last year of the DRAGUN (**D**etection, **R**etrieval, and **A**ugmented **G**eneration for **U**nderstanding **N**ews) track, which in its first year was called the Lateral Reading track [9]. The previous 2024 track had two tasks: generating questions to evaluate the trustworthiness of a given news article, and retrieving documents to answer those questions. Our analysis of the 2024 submissions revealed two key findings. First, questions generated by large language models (LLMs) showed little overlap with those created by TREC expert assessors. Second, many systems performed well at retrieving relevant documents from the specified corpus, suggesting that document retrieval was not the main challenge for developing systems to support people's assessment of news trustworthiness.

Building on these insights, we designed the TREC 2025 DRAGUN Track[1] with a more reader-focused approach. The 2025 track's primary task was to generate a well-sourced report that gives readers the context needed to assess a news article's trustworthiness themselves, not by labeling content as true or false. We continued the question generation task from 2024, as it remains challenging and is critical for guiding trustworthiness assessment, with slight modifications of submission requirements and a revamp of its evaluation to be consistent with the report evaluation. We made three major improvements based on lessons from the previous track:

- **New report generation task**: We replaced document retrieval with report generation, a retrieval-augmented generation (RAG) task. Participants had to create comprehensive

[1] `https://trec-dragun.github.io/`

Table 1: Example rubric question with short answers for a news article about the Epic Games vs. Apple lawsuit published on *The Verge*.

News Article
Title: Epic Games is suing Apple **URL:** `https://www.theverge.com/2020/8/13/21367963/`
Question 2: What is The Verge? **Importance:** Have to Know
Short Answers: 1. The Verge is a technology news website, located in New York City, and owned by Vox Media. *Reference:* `https://en.wikipedia.org/wiki/The_Verge` 2. It is a left-center website with high factual reporting and high credibility. *Reference:* `https://mediabiasfactcheck.com/the-verge/` 3. In the last five years, it has had no failed fact checks. *Reference:* `https://mediabiasfactcheck.com/the-verge/`

reports grounded in the MS MARCO V2.1 Segmented Corpus[2], the same corpus used by the TREC RAG Track[3]. This shared resource encouraged cross-track participation.

- **Starter-kit system**[4]: To lower the participation barrier, we released a complete baseline system with intermediate execution results at track launch [8]. This system uses a modular iterative multi-agent RAG pipeline. Using the starter kit, participants could quickly build their own systems by modifying or replacing individual components.
- **Enhanced evaluation rubrics**: Instead of having assessors write questions appropriate to support lateral reading in the 2024 track, we asked them to do their own research for a given news article to determine what they felt was important to know to understand its trustworthiness. Assessors could spend several hours doing this research using any existing online search engines and tools. Having determined what someone should know, assessors then created comprehensive rubrics to be used for the evaluation of the participants' generated questions and reports. These rubrics contain key questions paired with short answers that represent information genuinely useful for evaluating news articles. Table 1 shows an example question with its short answers. These rubrics also enable evaluation of systems using other web collections.

Participation increased substantially from the previous year. We received 37 runs from 10 teams for Task 1 (Question Generation) and 28 runs from 8 teams for Task 2 (Report Generation).

2 Tasks

The TREC 2025 DRAGUN Track consisted of two complementary tasks that participants could complete independently or together. Both tasks shared the same corpus and topics, with the same submission deadline. We designed these tasks to work together: Task 1 was to identify critical questions for a given news article, while Task 2 was to create a report that answers those questions. This design reflects our core philosophy: we help readers evaluate trustworthiness themselves by

[2] `https://trec-rag.github.io/annoucements/2025-rag25-corpus/#-ms-marco-v21-segmented-corpus`
[3] `https://trec-rag.github.io/`
[4] `https://github.com/trec-dragun/2025-starter-kit`

providing comprehensive context from a neutral perspective, rather than establishing an "absolute truth" and delivering a verdict.

2.1 Corpus

Both tasks used the MS MARCO V2.1 Segmented Corpus, the same as the TREC RAG Track (2024 and 2025). This corpus contains approximately 114 million segments created from 11 million web documents. Each segment is a sliding window of 10 sentences with a stride of 5 sentences. For Task 2, all generated reports had to cite specific segments from this corpus.

2.2 Topics

We selected 30 news articles from the MS MARCO V2.1 Document Corpus to serve as topics. These articles were chosen based on two criteria: they covered controversial issues from their publication period (2019–2021, close to the cut-off date when the corpus was collected) or contained content that would attract readers to seek additional context. Each article represents one topic for both tasks. The selected articles come from 28 different media sources with diverse political perspectives. According to Media Bias/Fact Check[5] ratings from May 2025, our topics include: 13 articles from left-leaning sources, 5 from right-leaning sources, 2 from neutral sources, 4 from pro-science sources, 2 from conspiracy-pseudoscience sources, and 4 from sources without established bias ratings.

2.3 Task 1: Question Generation

For each topic (target news article), participants needed to generate 10 critical questions that a thoughtful reader should investigate when assessing the article's trustworthiness, on aspects such as source bias, motivation, or alternative viewpoints. The generated questions were to be ranked from most to least important. Questions needed to meet the following requirements:

- A question's length could not exceed 300 characters.
- Questions should not be compound (e.g., *Who is X and when did Y happen?*). Each question should focus on a single topic.
- Given the context of the article, the questions should not be ambiguous or overly general. For example, "*Are there other sources corroborating the details presented in this article?*" is overly general.

Submissions could be either *manual*, i.e., involving human intervention to generate questions, such as hiring people to produce questions or manually selecting questions from candidate questions generated by algorithms, or other human involvement, or *automatic*, i.e., produced entirely by automatic systems without human input beyond the construction of the systems. Following standard TREC practice, teams submitting automatic runs needed to make a good-faith effort not to read or study the articles.

2.4 Task 2: Report Generation

Report generation is the core task of this track. For each news article, participants were to generate a 250-word well-attributed report that provides readers with essential background and context for evaluating trustworthiness. Report generation can be understood as a RAG task with a fixed query: "*What should I know about this article to better assess its trustworthiness?*", but with a varying

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

context (the news article). Each sentence of the report could have at most three references (i.e., segment IDs). Similar to Task 1, runs could be either automatic or manual.

3 Evaluation

During the participation period, TREC assessors created topic-specific rubrics after conducting open-web research on each article's trustworthiness, using any tools they preferred, e.g., web search. After runs were submitted, assessors used these rubrics to judge each system's questions and reports. We then computed overlap-based scores between each run and the rubric. The full assessment instructions are available in Appendix A.

3.1 Rubric Development

Each topic (news article) was assigned to one primary TREC assessor and two secondary TREC assessors. Each of the assessors conducted independent research on the trustworthiness of the news article using any tools and resources they liked. They created rubrics consisting of questions and answers that they thought a good report should cover to help readers determine the trustworthiness of the article. The primary assessor merged the rubrics from all three assessors into a single rubric. These rubrics were then examined by the organizers, and in a few cases, the answers were edited or split into multiple short answers for brevity. For example, we split one long answer "*Epic Games is a Cary, NC video game producer, founded by Tim Sweeney, while he was a student at the University of Maryland, in 1991. Their most popular game is Fortnite.*" into two shorter answers "*Epic Games is a Cary, NC video game producer, founded by Tim Sweeney in 1991.*" and "*Their most popular game is Fortnite.*" The final rubric contains a list of questions with short answers. Each question has an importance label:

- **Have to Know** (4 points): Core, critical questions. Knowing the answer is essential for judging the article's trustworthiness (it might change a reader's perception).
- **Good to Know** (2 points): Important contextual questions. Not absolutely critical, but answering them will increase a reader's confidence in their judgment.
- **Nice to Know** (1 point): Background or peripheral questions. These provide helpful context but are not crucial for most readers' trust decisions.

Table 1 shows an example of a rubric question with expected short answers.

3.2 Question Evaluation

The key idea of the scoring mechanism is to check how many rubric questions are covered by the participant-submitted questions. Due to budget constraints, the assessors only judged a subset of the question pairs (rubric question, participant question). For each rubric question, we used two models (`Qwen3-Embedding-8B`[6] and `Qwen3-Reranker-8B`[7]) to automatically select the most similar participant question. Specifically, we used the embedding model, `Qwen3-Embedding-8B`, to convert the rubric question and ten participant questions into vectors and picked the participant question that had the highest cosine similarity with the rubric question. Meanwhile, we concatenated the rubric question and each participant question as input to the reranker model, `Qwen3-Reranker-8B`, and picked the participant question with the highest reranking score.

[6] `https://huggingface.co/Qwen/Qwen3-Embedding-8B`
[7] `https://huggingface.co/Qwen/Qwen3-Reranker-8B`

For each rubric question, the assessor judged one (if both models picked the same most similar question) or two questions from the participant's question list, and assigned one of the following four similarity labels to the question pair:

- **Very Similar** (1 point): Questions may have different wording, but answering either question provides effectively the same information to the reader.
- **Similar** (0.5 points): Answering the questions will provide similar, but slightly different information to the reader.
- **Different** (0 points): The answer to each question will provide different information, with possibly some overlap, to the reader.
- **Very Different** (0 points): Answers to questions provide different information, with little to no overlap, to the reader.

For each run and each article, we first iterated through the list of assessor questions. For each assessor question, we checked if there were participant questions labeled as either "very similar" or "similar". If so, the run was rewarded with a score of (assessor question importance score $\times$ highest assessor similarity score). For example, regarding the rubric question in Table 1 that has an importance level of "have to know" (4 points), there is one participant question labeled as "very similar": "*Who owns The Verge, the outlet reporting Epic Games is suing Apple?*", and one participant question labeled as "similar": "*What political leaning is The Verge known for in coverage of large technology firms?*", from a run. For this rubric question, this run got $4 \times 1 = 4$, only rewarded by the "very similar" label. The same process was repeated for each of the remaining assessor questions, and then we summed up the scores and normalized the sum by the maximum possible score (the sum of all question importance scores) to get the run's final score $\in [0, 1]$ for this article.

Regarding the second requirement mentioned in Section 2.3, i.e., the exclusion of compound questions, this constraint was not explicitly enforced during the human assessment phase. To ensure adherence to task guidelines, we used an open-weight LLM (`gpt-oss-120b`[8]) to automatically identify and filter out compound questions. We validated this classification approach by manually labeling a stratified sample of 100 compound and 100 non-compound questions identified by the model. The classifier demonstrated high reliability, achieving a True Positive Rate of 0.989 and a False Positive Rate of 0.124. Consequently, we adopted this automated filtering mechanism for the final evaluation.

Quick qualitative analysis suggests this filtering approach was strict. Many questions flagged as compound combined highly correlated sub-inquiries (e.g., *"Who owns and funds Atlas Obscura, and how might that influence its food-origin narratives?"*). While these questions are semantically cohesive, they technically violate the single-topic constraint. To maintain fairness and strictly uphold the task specifications, we prioritized compliance and removed all questions flagged by the model.

We acknowledge a limitation in this post-hoc filtering approach. Because the selection of participant questions for human judgment occurred prior to the removal of compound questions, it is possible that a compound question was selected as the most similar match for a rubric question. If such a question were subsequently filtered out, the run would receive no credit for that specific rubric question, even if a valid, non-compound alternative existed but with a lower similarity score.

[8] `https://openai.com/index/introducing-gpt-oss/`

3.3 Report Evaluation

The key idea of the scoring mechanism is to check how many answers are covered by the report. For each report, the assessor assigned each answer with one of the following four labels:

- **Supports**: The report provides an answer to the question, consistent with the key elements of the rubric answer.
- **Partial**: The report provides an answer to the question that contains some but not all key elements of the rubric answer.
- **Contradicts**: The report contains information that contradicts the rubric answer. If the report supports or has partial support for the rubric answer, but it also contradicts the answer, then the *contradicts* label takes precedence.
- **None**: The report has no support or connection with the rubric answer. This is the default.

We produced two scores: supportive and contradictory. For the supportive score, we assigned "supports" a score of 1 and "partial" a score of 0.5. Then, for an answer, the report will get a score of (answer score / the number of answers from that question $\times$ question importance score). For example, if a report is labeled as "partial" for the Short Answer 1 in Table 1, it will get a score of $0.5 \ / \ 3 \times 4 = 0.67$. We summed up the scores for all answers, and then normalized the sum by the maximum possible score (the sum of all question importance scores) to get the final supportive score of the report. Thus, the supportive score is between 0 and 1, and the higher, the better. For the contradictory score, we assign "contradicts" a score of 1, while treating the other three labels as 0. We used the same computational method as the supportive score to get the final contradictory score of the report. Thus, the contradictory score is also between 0 and 1, but the lower, the better.

4 Submissions and Results

Our starter-kit system [8] uses a modular iterative multi-agent RAG pipeline in which LLM-driven agents cycle through query generation, segment retrieval (BM25 followed by neural reranking and LLM-based selection), and evidence evaluation until sufficient information is collected, after which the system produces both questions and reports. For Task 1, we also created two commercial baselines using Perplexity Deep Research (`organizer-t1-perplex`) and ChatGPT 5 Pro Deep Research (`organizer-t1-chatgpt`) by providing the same instructions given to TREC assessors along with the article text through each product's web interface. Neither baseline is reproducible, but they provide useful reference points. We briefly summarize the approaches of participating teams (alphabetically).

- CUET [5] built a LangChain-based multi-stage pipeline using open-weight LLMs (`QwQ-32B`, `Qwen3-14B`, `Mistral-Small-24B`, and `DeepSeek-R1-Qwen-32B`) with structured prompting that incorporates article metadata, controlled decoding, and LLM-based multi-criteria scoring for question evaluation. They participated in Task 1 only.
- Team cycraft built on the starter kit for both tasks, enhancing it by providing contrastive examples to the LLM. For question generation, they also added an LLM-based feedback reranking step where `GPT-4.1` generates feedback on the questions and reranks them accordingly. They used `GPT-4.1` for generation and Cursor Agent for system design and implementation.
- DUTH_XANTHI [2] explored mid-sized open-weight instruction-tuned LLMs (`Qwen2.5` variants, `Yi-1.5-9B`, and `Mistral-7B`) run entirely locally without fine-tuning or API access,

using zero-shot prompting with semantic filtering (TF-IDF combined with maximal marginal relevance) for question generation and BM25 retrieval via Pyserini for report generation.

- HLTCOE [4] submitted the CRUCIBLE system, which inverts the standard RAG pipeline by first automatically generating nuggets (question–answer pairs) from retrieved documents and then generating task outputs conditioned on those nuggets. For DRAGUN, they applied the nugget ideation stage with `Claude Sonnet 4` using a customized prompt borrowing from the track guidelines, and reused the starter kit's retrieval results for report generation.
- Team h2oloo participated in Task 1 only. Their 2025 approach adapts the prompting strategies they developed for the 2024 Lateral Reading Track. They use two prompting variants: an "original" (naive) prompt that directly asks the LLM to generate 10 questions, and a "stepwise" prompt that employs a chain-of-thought procedure: the LLM first lists the article's authors and sources, generates questions about their credibility and biases, identifies key claims, formulates accuracy questions, selects the 10 most significant, condenses any compound questions into simple ones, and resolves them for readability without article context. Both variants include a post-processing length check step. In 2025, they updated these prompts and switched to `GPT-5` and `Qwen3-30B-A3B-Thinking` as the backbone LLMs.
- SCIAI submitted two system descriptions. Their first approach, CITADEL [7], employs a hybrid retrieval pipeline (BM25 with synonym expansion, reciprocal rank fusion, and cross-encoder reranking) coupled with an agentic draft–critique loop in which separate `GPT-4.1` generator and evaluator agents iteratively refine reports. Their second approach [3] uses query expansion techniques and a `DeBERTa`-based regression model trained on 2024 Lateral Reading Track data to rank generated questions, with a cross-encoder for deduplication.
- TREMA-UNH [6] built upon the starter kit, replacing its LLM component with a local `Qwen 2.5:7B` model served via Ollama, and introduced an adversarial critique module in which the LLM generates balanced or aggressive critiques of articles to guide query generation and evidence retrieval.
- UR_trecking [1] used `GPT-4o-Nano` for question generation with semantic filtering and K-means clustering for diversity, combined with chain-of-thought query expansion, BM25 retrieval, monoT5 reranking, and domain-level trustworthiness scores for report generation.
- WaterlooClarke used a GPT-based approach with automatically generated feedback in the loop for question generation. For report generation, they generated the report first using open web search via ChatGPT, then grounded it in the MS MARCO corpus.

4.1 Task 1: Question Generation

Table 2 reports the average scores across 30 topics. All runs were declared *automatic*. The Perplexity baseline (`organizer-t1-perplex`) achieved the highest average score (0.354), while the second to the fourth runs share very close scores. Using a paired, two-tailed t-test across topics, `organizer-t1-perplex` is not significantly better than the runs ranked 2–5 ($p \geq 0.05$). The first run it significantly outperforms is our `organizer-gpt-oss-t1` (rank 6, $p < 0.05$).

Even the top-performing run was only able to cover less than half of the assessor questions. This suggests that systems need better alignment with fact-checkers' focus. In our rubric, "Have to Know" questions carry more weight (4 points), so aligning question generation with these critical aspects should improve scores dramatically. We also see small differences between neighboring runs and many non-significant pairwise comparisons. This reflects high topic variability: it is hard for a

Table 2: Average scores across 30 topics for Task 1 (Question Generation) runs.

Team	Run ID	Score
coordinators	organizer-t1-perplex	0.354
h2oloo	h2oloo_gpt5_orig	0.307
SCIAI	Team02_Task1	0.304
h2oloo	h2oloo_gpt5_step	0.294
cycraft	cursor-enhanced	0.264
coordinators	organizer-gpt-oss-t1	0.235
coordinators	organizer-t1-chatgpt	0.231
WaterlooClarke	feedbackintheloop	0.230
coordinators	dragun-organizers-starter-kit-task-1	0.226
cycraft	feedback-rerank	0.225
HLTCOE	cru-claude-chatty	0.218
CUET	CUET-QwQ-32B	0.215
DUTH_XANTHI	garamp_mistral_7b	0.199
CUET	CUET-qwen14B-v2	0.195
DUTH_XANTHI	garamp_yi15_9b	0.193
h2oloo	h2oloo_qw3-30b_step	0.190
CUET	CUET-DeepSeek-R1-Qwen-32B	0.184
CUET	CUET-Mistral-Small-24B	0.182
DUTH_XANTHI	garamp_qwen25_72b	0.177
CUET	CUET-qwen14B-v3	0.175
h2oloo	h2oloo_qw3-30b_orig	0.175
CUET	CUET-qwen14B-v5	0.165
CUET	CUET-qwen4B-v3	0.165
HLTCOE	cru-claude	0.163
DUTH_XANTHI	garamp_qwen25_7b_imp	0.158
CUET	CUET-qwen4B-v2	0.153
TREMA-UNH	SK_MI_2	0.151
CUET	CUET-qwen14B-v1	0.150
DUTH_XANTHI	garamp_qwen25_14b	0.149
UR_trecking	UR_IW_run_1	0.145
TREMA-UNH	SK_MI_1	0.141
TREMA-UNH	ConvF_all-t12_5	0.130
HLTCOE	cru-most_common	0.129
TREMA-UNH	SK_Critique_MI_5	0.127
TREMA-UNH	SK_ConvinceF_MI_2	0.120
TREMA-UNH	ConvF_all_MI_5	0.120
CUET	CUET-unsloth-Mistral-Small	0.081

system to be consistently better across diverse articles. As fact-checkers' focus depends on article type, there is no one-size-fits-all template for investigative questions.

4.2 Task 2: Report Generation

Table 3 shows the average supportive and contradictory scores. Because reports are limited to 250 words and must cite MS MARCO V2.1 segments, and because rubrics were built from live-web research (which may include information outside the corpus), scores far from the theoretical maximum of 1 are expected.

Our starter-kit system with `GPT-4.1` (`dragun-organizers-starter-kit-task-2`) achieved the highest average supportive score (0.230). Differences among the top four runs are small. Our

Table 3: Average supportive and contradictory scores across 30 topics for Task 2 (Report Generation) runs. Higher supportive and lower contradictory scores are better.

Team	Run ID	Supportive	Contradictory
coordinators	dragun-organizers-starter-kit-task-2	0.230	0.013
SCIAI	SCIAI_03_04_Eight	0.218	0.005
SCIAI	SCIAI_03_02_Three	0.212	0.014
SCIAI	SCIAI_03_03_Five	0.201	0.014
HLTCOE	cru-ablR_	0.173	0.009
HLTCOE	cru-confirm-ansR_	0.170	0.011
HLTCOE	cru-ablR-conf_	0.166	0.025
SCIAI	Team02_Run02_100SegmentsExpansion	0.158	0.013
HLTCOE	cru-clod-ablR-conf_	0.158	0.013
HLTCOE	cru-cloch-ablR-conf_	0.157	0.018
SCIAI	Team02_Run01_1000SegmentsExpansion	0.155	0.007
coordinators	organizer-gpt-oss-t2	0.150	0.018
SCIAI	Team02_Run03_100SegmentsNoExpansion	0.139	0.007
cycraft	cursor-report	0.132	0.008
SCIAI	Team01_Run01_Winner	0.125	0.005
TREMA-UNH	ConvF_all-t12_5_RG	0.123	0.009
TREMA-UNH	SK_MI_2_RG	0.110	0.008
TREMA-UNH	SK_Critique_MI_5_RG	0.098	0.014
WaterlooClarke	garag_rubric	0.097	0.018
TREMA-UNH	SK_ConvinceF_MI_2_RG	0.086	0.009
TREMA-UNH	ConvF_all_MI_5_RG	0.079	0.010
UR_trecking	UR_IW_run_1_task2	0.078	0.013
SCIAI	03_01_Baseline	0.067	0.003
DUTH_XANTHI	garamp_dragun_t2_q7b	0.005	0.002
DUTH_XANTHI	garamp_qwen25_14b_r4	0.005	0.002
DUTH_XANTHI	garamp_qwen25_3b_t2	0.005	0.002
DUTH_XANTHI	garamp_yi9b_t2_v1	0.005	0.002
DUTH_XANTHI	garamp_zephyr7b_t2	0.005	0.002

dragun-organizers-starter-kit-task-2 run is significantly better than the fifth-place cru-ablR_ at $p < 0.05$. All runs have low contradictory scores, showing relatively few contradictions against rubric answers. This indicates that factuality errors are not the primary challenge in this task. Several runs from the "DUTH_XANTHI" team often output the prompt text instead of a report, and therefore have both very low supportive and contradictory scores.

Similar to Task 1, topic effects are strong: differences between adjacent runs are small, and the top four runs are not significantly different from each other. We expect report quality to improve if systems first identify the most critical questions (Task 1) as fact-checkers and then focus the report on answering those questions.

5 Reuse of Track Resources

To make the track's evaluation reusable beyond the originally judged submissions, we have created an LLM-based AutoJudge system that can score new runs against the assessor-created rubrics without requiring further assessor effort [10]. In addition, we have collected and provided links to all the various track resources as well as the AutoJudge system via a GitHub repository[9].

[9] https://github.com/trec-dragun/resources

Resources include the 30 topics with their assessor-authored rubrics, all human judgments for both tasks, participant submissions, our starter-kit baseline system, and scoring scripts. While many of these resources are hosted by NIST, our repository provides a unified, documented view of them.

The AutoJudge uses few-shot prompting with `gpt-oss-120b`, mirroring the assessment instructions given to TREC assessors and including labeled examples from our organizer baseline runs. It outputs the same categorical labels as the human judging protocol, and scores are computed using the same scoring scripts. When validated against official human judgments, the AutoJudge preserves run-level rankings well, achieving Kendall's $\tau = 0.678$ for Task 1 and $\tau = 0.872$ for Task 2. Details of the AutoJudge design and validation are provided in a companion resource paper [10].

6 Conclusion

The DRAGUN track extended last year's Lateral Reading work by asking systems to produce well-attributed reports that could help readers judge the trustworthiness of online news. The track posits that AI systems should support human judgment, not replace it; they should surface broad and reliable context that helps readers make informed decisions. Our results show that current systems, including strong baselines, still miss a substantial portion of what expert assessors consider important. The main gap is not factuality, as contradictions are rare, but focus: systems need to target the key questions that matter most to readers' judgments. Future improvements should prioritize alignment with expert fact-checkers' priorities, especially the "Have to Know" items, and tighter coupling between question generation (Task 1) and report writing (Task 2).

Acknowledgments

This research received funding from Microsoft and the Natural Sciences and Engineering Research Council of Canada (NSERC) grant ALLRP/597573-24. We extend our appreciation to all participants who submitted work to the 2025 track. We are particularly grateful to Ian M. Soboroff and Hoa T. Dang at NIST for their coordination of the run assessment process, and to all TREC assessors who contributed their expertise to the evaluation.

References

[1] Ignacy Alwasiak, Kene Nnolim, Jaclyn Thi, Samy Ateia, Markus Bink, Gregor Donabauer, David Elsweiler, and Udo Kruschwitz. From Questions to Trust Reports: A LLM-IR Framework for the TREC 2025 DRAGUN Track. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[2] Georgios Arampatzis, Ioannis Maslaris, and Avi Arampatzis. LLM-Based Question Generation and Retrieval-Augmented Reporting for News Credibility. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[3] Jack Cheverton, Oliwia Majtyka, and Ting Liu. Intelligent News Comprehension through Query Expansion and LLM-Augmented Generation. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[4] Laura Dietz, Bryan Li, James Mayfield, Dawn Lawrie, Eugene Yang, and William Walden. HLTCOE Evaluation Team at TREC 2025: RAG, RAGTIME, DRAGUN, and BioGen. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[5] Adnan Faisal and Shiti Chowdhury. A LangChain-Based Framework for Investigative Question Generation Using Large Language Models. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[6] Naghmeh Farzi and Laura Dietz. TREMA-UNH at TREC 2025 DRAGUN Track: Iterative Multi-Agent Pipeline for News Verification via Adversarial Credibility Analysis with Local LLMs. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[7] Daniel Seredensky, Dylan Iddings, and Sharon G. Small. CITADEL — Citation-Driven Draft-Evaluate Loop. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[8] Dake Zhang. An Iterative Multi-agent RAG System for the TREC 2025 DRAGUN Track. In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

[9] Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. Overview of the TREC 2024 Lateral Reading Track. In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024)*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2024. URL `https://trec.nist.gov/pubs/trec33/papers/Overview_lateral.pdf`.

[10] Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. Resources for automated evaluation of assistive rag systems that help readers with news trustworthiness assessment, 2026. URL `https://arxiv.org/abs/2602.24277`.

A TREC 2025 DRAGUN Track Assessment Guidelines

Track Organizers: Dake Zhang, Mark Smucker, Charles Clarke
Track Website: TREC 2025 DRAGUN Track Guidelines

A.1 Overview

The **TREC 2025 DRAGUN Track** is designed to support readers in assessing the trustworthiness of online news articles. In this track, participants work on two tasks:

- ★ **Task 1: Question Generation.** Participants identify critical questions that a reader should consider when evaluating a news article's trustworthiness.
- ★ **Task 2: Report Generation.** Participants create a well-attributed, comprehensive report that provides relevant background and context, helping a reader make an informed judgment about the article's credibility. A good Task 2 report should address the important questions from Task 1.

As an **assessor**, you will **evaluate participants' questions and reports** to judge how well they help readers assess an article's trustworthiness. This involves three main responsibilities:

1. **Rubric Creation (Article Research & Criteria Development)**: For each assigned news article, you will conduct your own fact-checking and research on the article. Based on this investigation, you will develop a **rubric** – essentially **a set of key questions** or criteria that a high-quality report should address. Each question in your rubric will be accompanied by some **expected answers** with **supporting references**. This rubric represents the standard against which participant questions and reports will be evaluated.

2. **Question Evaluation**: You will assess a subset of questions submitted by participants. For each topic, you will be presented with some pairs of questions (one from your final rubric and one from the questions submitted by participants), and you will judge how similar the two questions are, using a 4-point scale (very similar, similar, different, or very different).

3. **Report Evaluation**: You will use your rubric to evaluate how well each submitted report covers the necessary information. In practice, you will check each report to see which of your rubric answers are covered by the report's content.

Neutral Point of View: *It is important to approach the assessment neutrally. Unlike traditional fact-checking that might conclude an article is simply "true" or "false", the DRAGUN track aims to help readers form their own judgments. Your rubrics should provide multi-source context from a neutral perspective, rather than asserting an "absolute truth". You are not expected to explicitly label an article as trustworthy or not; instead, you identify what information a well-rounded report should include for the reader's benefit.*

A.2 Example: A News Article and Its Rubric

To understand the expected output of the assessor's first task, consider the following example. Suppose you are given a news article titled "*Wildfire apocalypse, not as usual, the media's knee-jerk take on the Canadian wildfires was all wrong*" (an opinion piece by Steve Milloy, published June 12, 2023, in *The Spectator*). The article argues that media reports falsely blamed the Canadian wildfires on climate change.

Article URL: `https://thespectator.com/topic/wildfire-apocalypse-canada-climate-change/`

After reading the article and conducting research on its claims, source, and context, an assessor might create a rubric like this (individual assessors produce **unlabeled** questions; labels are added later by the primary assessor):

- **[Have to know]** Question 1: What should I know about the publisher of this article, The Spectator?
  - Answer 1: The Spectator is a politically conservative, right-leaning magazine.
    - Reference 1: `https://en.wikipedia.org/wiki/The_Spectator`
    - Reference 2: `https://www.allsides.com/news-source/spectator-world-media-bias`
    - Reference 3: `https://mediabiasfactcheck.com/the-spectator-usa/`
- **[Have to know]** Question 2: What should I know about the author of this article, Steve Milloy?
  - Answer 1: He is the "founder and editor of the blog JunkScience.com".
    - Reference 1: `https://en.wikipedia.org/wiki/Steven_Milloy`

        - Reference 2: https://junkscience.com/who-is-steve-milloy/
    - Answer 2: He has "close and long-standing financial ties to oil companies".
        - Reference 1: https://en.wikipedia.org/wiki/Steven_Milloy
        - Reference 2: https://www.motherjones.com/environment/2005/05/some-it-hot/
    - Answer 3: He denies the scientific consensus on climate change.
        - Reference 1: https://en.wikipedia.org/wiki/Steven_Milloy
        - Reference 2: https://www.eenews.net/articles/steve-milloy-doesnt-like-climate-bedwetters/
- **[Have to know]** Question 3: What is the Environmental Protection Agency (EPA)?
    - Answer 1: It is "an independent agency of the United States government tasked with environmental protection matters".
        - Reference 1: https://en.wikipedia.org/wiki/United_States_Environmental_Protection_Agency
        - Reference 2: https://www.epa.gov/aboutepa/our-mission-and-what-we-do
    - Answer 2: It "conducts environmental assessment, research, and education".
        - Reference 1: https://en.wikipedia.org/wiki/United_States_Environmental_Protection_Agency
- **[Have to know]** Question 4: What is the relationship between wildfires and climate change, according to research studies?
    - Answer 1: "Research shows that changes in climate create warmer, drier conditions, leading to longer and more active fire seasons."
        - Reference 1: https://www.noaa.gov/noaa-wildfire/wildfire-climate-connection
        - Reference 2: https://science2017.globalchange.gov/chapter/8/
    - Answer 2: "Wildfire activity appears strongly associated with warming and earlier spring snowmelt."
        - Reference 1: https://royalsocietypublishing.org/doi/10.1098/rstb.2015.0178
- **[Have to know]** Question 5: The article claims that "Canadian government data show that wildfires in Canada have been overall declining since 1980". Is this claim accurate?
    - Answer 1: "There are fewer fires, but an increase in area burned and number of people displaced."
        - Reference 1: https://www.cbc.ca/news/climate/canada-wildfire-data-change-1.6854186
- **[Good to know]** Question 6: What is the Energy and Environment Legal Institute, with which Steve Milloy is affiliated?
    - Answer 1: It is a "right-of-center nonprofit legal advocacy group".
        - Reference 1: https://www.influencewatch.org/non-profit/energy-and-environment-legal-institute/
    - Answer 2: It "opposes environmentalist legislation".
        - Reference 1: https://www.influencewatch.org/non-profit/energy-and-environment-legal-institute/
    - Answer 3: It is one of the groups funded by coal companies.

        - Reference 1: https://www.theguardian.com/environment/2016/jun/13/peabody-energy-coal-mining-climate-change-denial-funding

- **[Good to know]** Question 7: What are the known health effects of PM2.5 (fine particulate matter), according to research?
    - Answer 1: "Breathing in unhealthy levels of PM2.5 can increase the risk of health problems like heart disease, asthma, and low birth weight."
        - Reference 1: https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm
    - Answer 2: "Increased daily mortality is specifically associated with particle mass constituents found in the aerodynamic diameter size range under 2.5 microns."
        - Reference 1: https://pubmed.ncbi.nlm.nih.gov/8875828/

- **[Good to know]** Question 8: Is the number of wildfires declining around the world?
    - Answer 1: "Extreme wildfires have become more frequent, more intense, and larger."
        - Reference 1: https://www.wri.org/insights/global-trends-forest-fires
        - Reference 2: https://science.nasa.gov/earth/explore/wildfires-and-climate-change/
    - Answer 2: "The decline in global average area burned has indeed been misused to support false claims numerous times."
        - Reference 1: https://royalsociety.org/blog/2020/10/global-trends-wildfire/

- **[Nice to know]** Question 9: Are springtime wildfires common in Canada?
    - Answer 1: In Canada, "most wildfires occur between April and September".
        - Reference 1: https://www.redcross.ca/how-we-help/emergencies-and-disasters-in-canada/types-of-emergencies/wildfires/wildfires-information-facts
    - Answer 2: "A spring wildfire season is common in British Columbia."
        - Reference 1: https://blog.gov.bc.ca/bcwildfire/spring-2024-seasonal-outlook/

- **[Nice to know]** Question 10: What is the relationship between PM2.5 and asthma?
    - Answer 1: "Long-term exposure to PM2.5 has significant adverse effects on childhood and adult asthma."
        - Reference 1: https://www.sciencedirect.com/science/article/pii/S2590332224004871
    - Answer 2: Air pollution, including PM2.5, can cause asthma and other lung diseases.
        - Reference 1: https://aafa.org/asthma/asthma-triggers-causes/air-pollution-smog-asthma/
    - Answer 3: A study in 2021 "found insufficient evidence to determine the effect of PM2.5 on asthma in the indoor home environment."
        - Reference 1: https://www.sciencedirect.com/science/article/pii/S0013935121009257
    - Answer 4: A study in 2024 found "positive correlation between PM2.5 concentration and the cumulative incidence of asthma, with a lag of 0–7 days".
        - Reference 1: https://aaqr.org/articles/aaqr-23-08-oa-0195

How to read this rubric: In the above example, the assessor identified ten questions that a thorough report on the article should answer. Each question is labeled by importance:

- **Have to know:** Core, critical questions. Knowing the answer is essential for judging the article's trustworthiness (it might change a reader's perception).
- **Good to know:** Important contextual questions. Not absolutely critical, but answering them will increase a reader's confidence in their judgment.
- **Nice to know:** Background or peripheral questions. These provide helpful context but are not crucial for most readers' trust decisions.

For each question, the assessor provided **expected short answers** – concise facts or findings that answer the question – along with **references** (links to sources) supporting each answer. A participant's report that addresses all the "Have to know" and most "Good to know" questions with credible references would be considered very strong. This example illustrates the format and level of detail expected in the rubrics you will create.

A.3 Developing the Rubric for an Article

Your first major task is to investigate your assigned news article and **create a rubric** based on your findings. In essence, you will be performing your own fact-checking research (like writing a mini-report for yourself) and then translating that into a set of Q&A criteria. (There is an optional online self-paced training, in Section A.6, that gives you hints on what aspects you can investigate for a news article.) Follow these steps to develop your rubric:

1. **Read the Article and Conduct Your Research.** Begin by reading the news article (topic) you've been assigned. During your reading, you are encouraged to conduct web searches using search engines of your preference to help you understand claims or statements, the context in which it was written (date, author, publication), and any references or sources cited in the article itself. **Keep track of your browsing history, either by keeping those tabs open or taking notes.** This investigative stage is crucial – you are essentially gathering the information that a well-informed reader should know before trusting the article. Key things to research include:

    a. **Reputation and bias** of the publisher (e.g., what do we know about the website or outlet?).

    b. **Background** of the author (expertise, affiliations, any potential agenda).

    c. **Veracity of key claims or statistics** in the article (checking if they are supported or refuted by reliable data).

    d. **Broader context**: For example, if the article is about a scientific topic or an event, find what scientific research or authoritative reports say about it.

2. **Identify the Key Issues and Questions.** Based on your research, determine what questions a comprehensive report should answer to cover all the important aspects of the article's trustworthiness. Think about it this way: after doing your research, you now know a lot about the article's claims and context – what are the most important points a reader should be aware of? For example, if the article makes a scientific claim, one question might be "What do scientific studies say about [claim]?" If the article's author has a clear bias or affiliation, a question might be, "What bias does [author name] have when reporting [subject]?" Make a

list of these potential questions. Aim for **about 5 to 10 questions** per article as a guideline (there's no strict rule – some articles might need a bit more or fewer). Ensure the questions **collectively cover all major facets** of evaluating the article's trustworthiness.

3. **Formulate Clear, Focused Questions.** Now, refine the wording of each question on your list. Each question in your rubric should be clear and **focused on a single aspect**. **Avoid compound questions** that ask about multiple things at once (e.g., "Who is the author and what is their affiliation?"). Also, **avoid overly broad or vague questions** like "Is the article credible?" If a question seems too general, try tying it to a specific claim or element in the article. For example, rather than a broad "Are there other sources corroborating the article's claims?", you might ask a more pointed question about a particular claim: "Have other sources corroborated the article's claim that wildfire frequency is declining?" Focus on what a reader needs to know versus what might simply be curiosities.

4. **Research and Draft Expected Answers for Each Question.**

    a. For every question you include in the rubric, write one or more concise **answers** that directly address that question based on your research. An answer is usually one sentence that captures the key fact or insight the reader should learn. It should be factual and to the point. You may phrase it in your own words or quote directly from a source – if you quote, use quotation marks. The answer should be something you could imagine appearing in a well-researched report.

    b. Every answer must be backed up with **at least one reference (URL)** supporting that answer. **Each reference (URL to a web page) should independently verify the answer.** Use reliable, credible sources – prioritize things like established news outlets, academic papers, authoritative reports, or well-regarded reference sites. It's acceptable to use the same reference for multiple answers if it contains information relevant to each, but avoid overly relying on a single source for everything. Also, use **English-language text-based sources** (i.e., the answer can be derived from the text content of the web page, not pictures, audio, or videos) and sources that are likely to remain accessible (for example, a static article instead of a temporary social media post). If you find yourself needing to combine information from multiple references to answer one question, consider splitting into multiple answers, each with its own reference, so that each answer is straightforward and well-supported.

5. **Review and Refine the Rubric.** *Skip importance labels for now; they will be assigned later by the primary assessor during consolidation.* Finally, read through your entire rubric (questions, answers, references) and check for completeness and clarity:

    a. Do the questions cover all major concerns about the article's trustworthiness? Imagine someone reads the participant's report – after answering all these questions, would they feel well-equipped to decide if the original article is trustworthy?

    b. Are the questions phrased clearly, without bias or implying a judgment? (We want neutral questions. For instance, instead of "Why is this article wrong about climate change?", a neutral phrasing would be "What do scientific sources say about the cause of climate change?")

    c. Are the answers factual, concise, and fully supported by the cited references? Double-check that each reference indeed backs up the answer. If the connection isn't obvious, either clarify the answer or choose a more direct source.

Collaboration and Consolidation: Each article (topic) in this track will be assigned to **three assessors** – one **primary assessor** and two **secondary assessors**. All assessors work **independently** on the above steps to create their own unlabeled rubric (5–10 questions each) for the article. After that, the **primary assessor** will gather the rubrics from the secondaries and **fuse** them into **one final rubric of no more than 10 questions**. Use the following guidelines when assigning importance labels (i.e., how essential that question is for judging the article's trustworthiness), but the primary assessor's judgment prevails:

- Have to know: This question addresses a core aspect of trustworthiness. The answer is critical for a reader to make an informed judgment.
- Good to know: This question covers important context that strengthens a reader's confidence in their judgment, though it may not be absolutely make-or-break.
- Nice to know: This question provides useful background or additional context that is helpful but not essential. These are more like bonus information that enriches understanding.

If a question appears in all three rubrics, it's highly likely to be a "have to know" question; if it appears in two, it is probably a "good to know"; and if it appears in only one, it is likely "nice to know." Consider also the perspective of an average reader: What information must they have to evaluate the article, and what information would merely be helpful or interesting? Use these labels to prioritize the questions accordingly.

The **core goal** here is not to dream up tricky questions, but to leverage thorough research to pinpoint what information **needs to be in a good report**. In effect, you first create a research-based *mini-report* for yourself on the article, and then extract from it the questions and answers that will form your rubric. By focusing on factual findings and key trustworthiness factors, you ensure your rubric is grounded in evidence and directly tied to the track's objectives. The final rubric will then be used in question evaluation and report evaluation.

A.4 Question Evaluation

For each article (topic), you (primary assessor) will assess a subset of questions submitted by participants. You will be presented with some pairs of questions (one from the final rubric and one from participants), and you will judge how similar the two questions are, using a 4-point scale (very similar, similar, different, or very different).

- **Very Similar**: Questions may have different wording, but answering either question provides effectively the same information to the reader. The two questions are the same exact question or two different ways of asking the same thing.
- **Similar**: Answering the questions will provide similar, but slightly different information to the reader. You can think of similar questions as attempts to interrogate the article and get to similar conclusions or information. For example, these two questions are similar because they are getting to the idea of LA Times credibility, but one is more specific:
  - Is the LA Times a credible newspaper?
  - Does the LA Times have a reputation for reporting fairly on criminal cases?
- **Different**: The answer to each question will provide different information, with possibly some overlap, to the reader. For example, you could have two questions about the same person, but asking different questions about that person, and these are different questions because they are not aiming to find out the same thing.

- **Very Different**: Answers to questions provide different information, with little to no overlap, to the reader. The two questions are asking different things.

Interpret each question in the context of the target news article to help you make sense of what the question is asking. When comparing against the rubric question, judge each participant question **independently** of the other participant questions. There may be grey areas where you might disagree with other assessors about the particular judgment. That's OK, but be consistent in how you apply your criteria for judging.

The assessment tool for the question assessment task is at: [URL]. Each of you should have already received an email with your username and password.

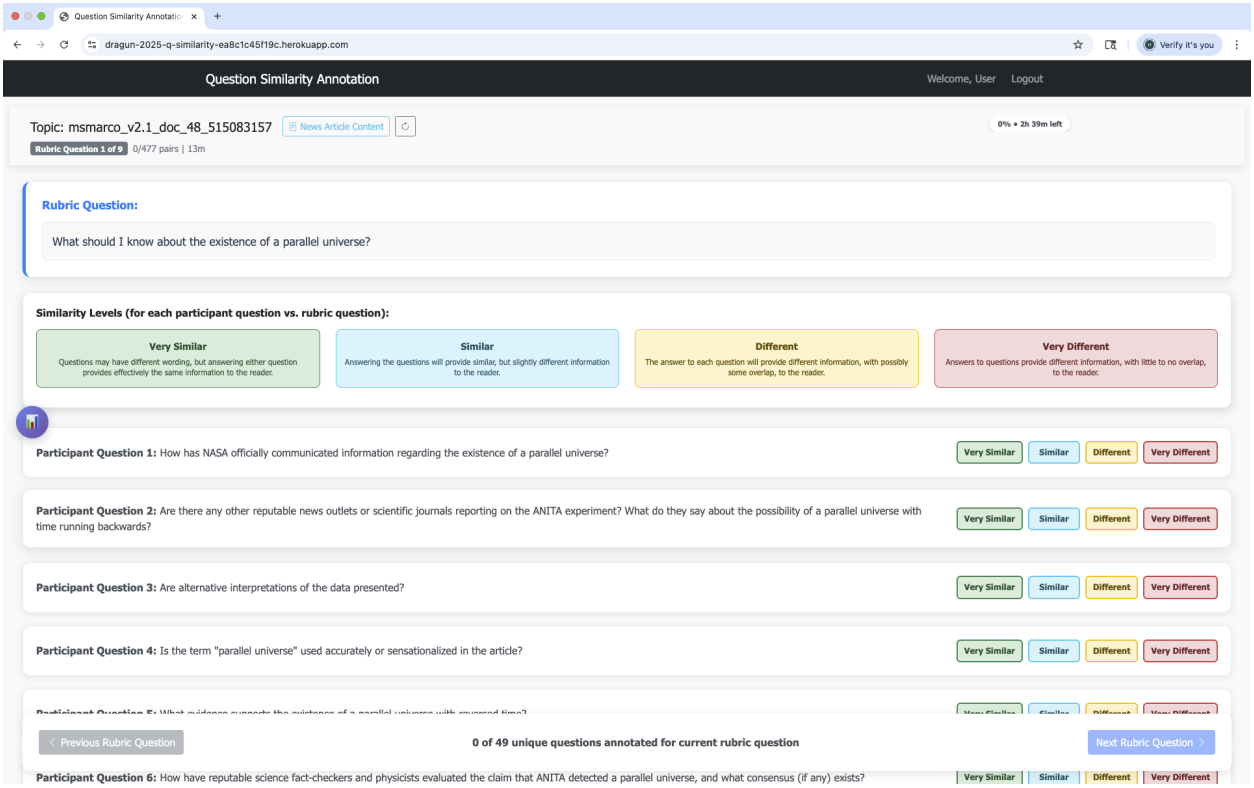

Figure 1: Question evaluation web interface.

How to use the web interface (shown in Figure 1)?

- The document ID of the target news article for the current topic is at the top of the window. Clicking on "News Article Content" will pop up another browser window/tab and display the text of the news article.
- The "Rubric Question" is the assessor question that you'll be comparing against the "Participant Question".
- The definition is displayed for each of the 4 judgments for question similarity (Very Similar, Similar, Different, Very Different)
- The "Participant Question" is the question that you're judging and comparing against the "Rubric Question". Click on one of the judgment buttons to the right of the Participant Question; You can change your judgments, and only the latest judgment for each question pair will be stored.

- **You will need to assess all question pairs for all the rubric questions for the topic (up to 10 rubric questions per topic).** The "Previous Rubric Question" and "Next Rubric Question" buttons at the bottom corners of the window let you switch between rubric questions for the topic. However, the "Next Rubric Question" button is enabled only after all question pairs have been assessed for the current rubric question.
- Clicking on the multi-colored floating symbol at the center-left of the window will display all of your assigned topics and allow you to switch between topics. There are progress bars that display the number of question pairs assessed and the estimated time remaining to finish the topic. The initial estimation is not accurate. It will become more accurate when you've judged more questions.

A.5 Report Evaluation

Overview: This section outlines how to evaluate Task 2 reports using the consolidated rubric. The goal is to determine how well each submitted report covers the necessary information for judging the news article's trustworthiness, according to your rubric. You will read the report, and then for each expected answer in the rubric, decide if the report **supports** the answer (consistent with the answer), **partially supports** the answer, has **contradictions** with the answer (inconsistent with the answer), or has **no support** or connection to the answer. To support an answer, the report should be consistent with the answer. When we say the report is "consistent with the answer," we don't mean it has to use the same words. We mean that the information in the report should agree with and not contradict what the answer says.

Another way to explain this is that for the DRAGUN track, the answers are effectively exemplars – good examples of the kind of information we want to see in the report – so you should check whether the report lines up with the meaning of the exemplar, not whether it copies it exactly. A report's answer is consistent if it says the same thing in different words, or provides information that fits with the exemplar answer. The report's answer is not consistent if it says something that disagrees with or directly contradicts the exemplar answer.

Evaluation Procedure:

1. **Use the Consolidated Rubric as a Checklist**: You will use the consolidated rubric (which contains the key questions and answers) as the basis for checking each report's content. The organizers made some changes when extracting your rubrics from your Google documents to the assessment tool, mostly fixing typos, reranking them based on the importance labels, and splitting long answers into smaller ones. Please check your rubrics shown on the assessment tool and let the organizers know if you feel some changes are wrong.
2. **Read the report and judge it against the answers for each question in the rubric.** For **each question** in the rubric and its expected answers, select one of the following judgments:
    - **Supports:** The report provides an answer to the question that is consistent with the key elements of the rubric answer.
    - **Partial:** The report provides an answer to the question that contains some but not all of the key elements of the rubric answer.
    - **Contradicts:** The report contains information that contradicts the rubric answer. If the report *supports* or has *partial* support for the answer, but it also contradicts the answer, then you should select *contradicts*.

- **None:** The report has no support or connection with the rubric answer. This is the default.

3. **Repeat for All Reports**: Repeat the above process for each report for the same article, evaluating it in the same manner. Once all reports for that article are done, proceed to your next assigned article.

Example: Labeling a Report

Now consider a very short report (one sentence) for the news article above:

> *"The Spectator article by Steve Milloy, a commentator known for rejecting the scientific consensus on climate change, argues that the Canadian wildfires and resulting smoke were a routine occurrence misattributed to climate change and overhyped by the media."*

Below is a rubric question with answers from the example above:

[Have to know] Question 2: What should I know about the author of this article, Steve Milloy?

- Answer 1: He is the "founder and editor of the blog JunkScience.com". [None]
- Answer 2: He has "close and long-standing financial ties to oil companies". [None]
- Answer 3: He denies the scientific consensus on climate change. [Supports]

This report **directly addresses** Answer 3, as it notes Milloy is known for rejecting the climate change consensus. It does **not** mention Milloy's blog or his ties to oil companies (Answers 1 and 2). Therefore, the assessor would mark **Answer 3 as "Supports"** for this report, and leave Answers 1 and 2 as **"None"**. All other questions' answers would remain "None" as well, since this report is only relevant to Question 2. The assessor should click the "Supports" option for Answer 3, and leave the default (None) for the others.

A.6 Online Self-Paced Training (Optional but Recommended)

Before or during your work on rubric creation, it's highly recommended to familiarize yourself with the verification skills that will be invaluable for this task. If you are already experienced in digital fact-checking and lateral reading, you may skim this section. Otherwise, consider completing the CTRL-F verification skills training, an online self-paced course developed by CIVIX Canada, which covers core techniques for assessing what information to trust online. The training takes approximately 2–3 hours. Use the checklist below to guide you through:

1. **Home**: Navigate to the CTRL-F student home page: `https://ctrl-f.ca/en/student/home/`.
   ☐ Read the page content.
   ☐ Watch the embedded video: Intro to Verification Skills — CTRL-F.
   ☐ Click the "Begin" button at the bottom of the page to start the training. This will take you to the "Why Verify?" tab.

2. **Why Verify?**: `https://ctrl-f.ca/en/student/why-verify/`
   ☐ Read the page content.
   ☐ Play the "FakeOut" game by clicking "Play now".
   ☐ Watch the embedded video: CIVIX Explains: Information Pollution.
   ☐ Click "Next" to continue to the "Source" tab.

3. **Source**: `https://ctrl-f.ca/en/student/source/`
   ☐ Read the page content.
   ☐ Watch the embedded videos:
      ☐ Investigate the Source — CTRL-F
      ☐ Skill: Just Add Wikipedia — CTRL-F
      ☐ Skill: Advanced Wikipedia – Bias & Agenda
      ☐ Why Use Wikipedia? (supplemental)
      ☐ Tips and Tricks for Using Wikipedia (supplemental)
      ☐ Evaluating Expertise — CTRL-F
      ☐ CIVIX Explains: Persuasive Sources
   ☐ Go through the three examples in the "Test your skills" section.
   ☐ Click "Learn the next skill" to continue to the "Claim" tab.

4. **Claim**: `https://ctrl-f.ca/en/student/claim/`
   ☐ Read the page content.
   ☐ Watch the embedded videos:
      ☐ Check the Claim — CTRL-F
      ☐ Skill: Check Other Sources — CTRL-F
      ☐ Skill: Advanced Claim Check — CTRL-F
   ☐ Go through the three examples in the "Test your skills" section.
   ☐ Click "Learn the next skill" to continue to the "Trace" tab.

5. **Trace**: `https://ctrl-f.ca/en/student/trace/`
   ☐ Read the page content.
   ☐ Watch the embedded videos:
      ☐ Trace the Information — CTRL-F
      ☐ Skill: Click Through & Find — CTRL-F

      (You may skip the video "Skill: Search the History of an Image," as our focus is on textual news content.)

   ☐ Go through the first example in the "Test your skills" section *(the remaining examples involve image history and can be skipped).*

6. Congratulations! You have completed the training!

A.7 Final Words

By following this guide, you will produce a detailed rubric that encapsulates what a trustworthy analysis of the news article should include, and you will be prepared to evaluate the participants' questions and reports with confidence and consistency. Thank you for your effort in this assessment process – your expertise is crucial to ensuring that the DRAGUN Track results in meaningful insights on how to assess news trustworthiness. Good luck with your assessments!