Resources for Automated Evaluation of Assistive RAG Systems that Help Readers with News Trustworthiness Assessment

Dake Zhang
University of Waterloo
Waterloo, Ontario, Canada

Mark D. Smucker
University of Waterloo
Waterloo, Ontario, Canada

Charles L. A. Clarke
University of Waterloo
Waterloo, Ontario, Canada

Abstract

Many readers today struggle to assess the trustworthiness of online news because reliable reporting coexists with misinformation. The TREC 2025 DRAGUN (**D**etection, **R**etrieval, and **A**ugmented **G**eneration for **U**nderstanding **N**ews) Track provided a venue for researchers to develop and evaluate assistive RAG systems that support readers' news trustworthiness assessment by producing reader-oriented, well-attributed reports. As the organizers of the DRAGUN track, we describe the resources that we have newly developed to allow for the reuse of the track's tasks. The track had two tasks: (Task 1) Question Generation, producing 10 ranked investigative questions; and (Task 2, the main task) Report Generation, producing a 250-word report grounded in the MS MARCO V2.1 Segmented Corpus. As part of the track's evaluation, we had TREC assessors create importance-weighted rubrics of questions with expected short answers for 30 different news articles. These rubrics represent the information that assessors believe is important for readers to assess an article's trustworthiness. The assessors then used their rubrics to manually judge the participating teams' submitted runs. To make these tasks and their rubrics reusable, we have created an automated process to judge runs not part of the original assessing. We show that our AutoJudge ranks existing runs well compared to the TREC human-assessed evaluation (Kendall's $\tau = 0.678$ for Task 1 and $\tau = 0.872$ for Task 2). These resources enable both the evaluation of RAG systems for assistive news trustworthiness assessment and, with the human evaluation as a benchmark, research on improving automated RAG evaluation.

CCS Concepts

• **Information systems** → **Retrieval tasks and goals**; **Evaluation of retrieval results**; **Test collections**; • **Computing methodologies** → **Natural language generation**.

Keywords

Text REtrieval Conference; Test Collection; Evaluation; Retrieval-Augmented Generation; News Trustworthiness

ACM Reference Format:
Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. 2026. Resources for Automated Evaluation of Assistive RAG Systems that Help Readers with News Trustworthiness Assessment. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3805712.3808624

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808624

1 Introduction

Retrieval-Augmented Generation (RAG) has made it practical to build language-model assistants that synthesize information and cite supporting sources [12]. This shift is especially relevant for tasks where users do not merely want a short answer, but instead need help assembling context and corroborating details across sources. As RAG systems move from short answers to structured reports, evaluation becomes a central bottleneck. Human assessment remains the most faithful way to compare systems, yet it is expensive and difficult to reuse when new systems appear. Meanwhile, standard automatic metrics often fail to reflect whether a generated report actually covers the information a careful reader needs and whether it does so without contradictory claims [10, 16].

This paper focuses on the trustworthiness assessment of online news, a setting where the evaluation problem is both socially important and technically subtle. Online news coexists with misinformation and low-quality reporting, and false or misleading content can spread quickly and widely [1, 11, 27]. Such exposure has been linked to shifts in trust, polarization, and other downstream societal harms [20, 25]. A growing literature therefore studies interventions and tools that help users slow down, reflect, and better distinguish reliable reporting from deceptive or unsubstantiated content before acting on it [3, 7, 14]. Yet for many readers, the limiting factor is not access to information but the ability to evaluate it effectively under time constraints, limited domain knowledge, and persuasive presentation tactics [18, 22].

A key insight from research on digital literacy is that expert fact-checkers behave differently from typical readers. Novices often read *vertically*, staying on the page and relying on on-page cues. In contrast, professional fact-checkers and skilled readers practice *lateral reading*: they quickly leave the page to investigate the publisher, trace claims to primary sources, and compare coverage across outlets [17, 29]. This workflow improves accuracy but requires knowing what to check and efficiently finding and weighing evidence. RAG systems have the potential to assist readers in this news trustworthiness assessment by helping them realize what questions matter for a particular article and providing evidence-grounded context that readers can directly use or inspect further.

Assessing an article's trustworthiness goes beyond checking a set of explicit, well-formed claims. Most prior automated fact-checking research has framed the problem as a pipeline of claim identification, evidence retrieval, and veracity prediction [8, 19]. Benchmarks such as FEVER [24] and LIAR [28] have been instrumental in advancing research on evidence retrieval and entailment-style verification. However, news reporting may be misleading due

to selective omission, missing context, or questionable sourcing that are not reducible to a single proposition. From the reader's perspective, the core task is closer to determining what to investigate and what context they should know to better assess an article's trustworthiness. This motivates evaluation settings that reward systems for surfacing the most important investigative angles and for producing to-the-point, multi-source, and well-supported reports.

The TREC 2025 DRAGUN (**D**etection, **R**etrieval, and **A**ugmented **G**eneration for **U**nderstanding **N**ews) Track was designed as an end-to-end benchmark for this assistive setting [31]. We defined two complementary tasks over a fixed retrieval corpus (MS MARCO V2.1 Segmented Corpus): (Task 1) generating a ranked list of ten investigative questions that a careful reader should ask of the target news article, and (Task 2, the main task) generating a 250-word report that covers what an informed reader should know to assess its trustworthiness, grounded in retrieved corpus segments with explicit attribution. The tasks were designed with reader utility in mind: rather than asking systems to output a single *true* or *false* label, DRAGUN requires systems to produce artifacts that support lateral reading by guiding investigation and summarizing corroborating context with verifiable citations.

For evaluation, TREC assessors conducted open-web research and produced importance-weighted rubrics consisting of focused questions with one or more expected short answers, each supported by reference URLs (one example shown in Table 1). These rubrics specify, for each news article, what an informed reader should know to assess its trustworthiness, and submitted runs were then judged against these rubrics. This rubric-first design helps the benchmark reflect expert investigative priorities (including angles that submitted systems may miss) and makes scoring interpretable: systems are rewarded for covering high-importance rubric items and penalized when they contradict rubric answers.

DRAGUN builds on the tradition of nugget-based evaluation [13, 26] in that rubric questions and their expected short answers play the role of weighted content units, and the evaluation reduces to judging whether a system's questions and report cover, partially cover, or contradict those units. To automate nugget-based evaluation, previous work such as Nuggeteer [15] relied on n-gram overlap and corpus-derived term weights to approximate nugget assignments, which was appropriate when robust semantic inference models were not available. In this paper, we instead leveraged advanced Large Language Models (LLMs) as rubric judges, aligning with recent evidence that LLM-based judges can preserve system rankings and approximate human preferences when appropriately prompted and calibrated [21, 23, 33]. This allows DRAGUN's rubric-based evaluation to scale beyond the original judged submissions.

Accordingly, this resource paper makes DRAGUN reusable by releasing not only the topics, rubrics, and human judgments, but also an automated LLM-based AutoJudge that can score additional runs against the released rubrics. Our AutoJudge mirrors the human judging protocol using few-shot prompting and produces the same categorical labels used in manual assessment. When validated against official human judgments, the resulting run-level rankings closely match the human-based rankings (Kendall's $\tau = 0.678$ for Task 1 and $\tau = 0.872$ for Task 2). Taken together, DRAGUN's released artifacts[1] enable both (1) benchmarking new assistive RAG systems for news trustworthiness assessment and (2) studying automated evaluation itself, using the human labels and rankings as a reference point:

(1) A rubric-based benchmark for lateral-reading-style assistance, including 30 news articles and assessor-authored, importance-weighted rubrics of investigative questions with expected short answers and supporting URLs.
(2) Human judgments and scoring code for both question generation and report generation, enabling reproducible evaluation and analysis of system behavior.
(3) An LLM-based AutoJudge that scales rubric-based evaluation to new runs, empirically preserving the official human ranking while keeping the assessments interpretable.

The remainder of this paper describes the DRAGUN tasks (Section 2), the human rubric construction and assessment procedures (Section 3), the AutoJudge design and validation (Section 4), and the released resources and intended reuse scenarios (Sections 5-6).

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/

2 Tasks

The TREC 2025 DRAGUN Track consisted of two complementary tasks; participants could complete either or both. 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 addresses those questions. This design reflects our core philosophy: we help readers evaluate trustworthiness themselves by providing comprehensive context rather than labeling content as true or false. We created a website to communicate task descriptions and submission requirements to participants: https://trec-dragun.github.io/.

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

2.1 Corpus and Topics

Both tasks used the MS MARCO V2.1 Segmented Corpus[2], which contains approximately 114 million segments derived from 11 million web documents. Each segment is a sliding window of 10 sentences with a stride of 5 sentences. For Task 2, generated reports had to cite relevant segments from this corpus.

We selected 30 news articles from the MS MARCO V2.1 Document Corpus (before segmentation) to serve as *topics*. These articles were chosen based on two criteria: they cover controversial issues from their publication period (2019-2021, close to the cut-off date when the corpus was collected) or contain content that would attract readers to seek additional context. The selected articles are from 28 different media sources with diverse political perspectives. According to Media Bias/Fact Check[3] ratings in 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.

2.2 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, regarding 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.

2.3 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 context, i.e., the news article. Each sentence of the report could have at most three references (i.e., segment IDs).

3 Human Assessment

The track's evaluation of submitted runs was completed by TREC assessors. The overall process was that during the track participation period, 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 the submitted questions and reports. We then computed scores based on their judgments. The remainder of this section provides the details of the rubric construction, question assessment, and report assessment processes.

3.1 Rubric Construction

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 deemed appropriate. 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 then merged the rubrics from all three assessors into a single final rubric (capped at ten questions). The final rubric contains a list of questions, each with one or more short answers. Table 1 shows an example rubric question with expected short answers. Each question also 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.

To standardize rubric quality, we provided assessors with detailed assessment guidelines emphasizing neutral, reader-oriented fact-checking via open-web research and lateral reading. Concretely, assessors were instructed to investigate (1) the publisher's reputation and potential bias, (2) the author's background, expertise, affiliations, and possible agenda, (3) the veracity of salient claims or statistics made in the article, and (4) the broader context from authoritative reports or research when relevant. Rubric questions were expected to be focused on a single aspect and phrased without implying an overall verdict on whether the article is true or false. For each rubric question, assessors wrote one or more concise short answers, and every short answer was required to be backed by at least one supporting reference URL. We then examined the written rubric answers, and for a few answers that were too long, we shortened them or broke them into multiple short answers. The guidelines encouraged using credible, English-language, text-based sources that are likely to remain accessible, and not over-relying on a single source.

3.2 Question Assessment

The key idea of question assessment is to check how many rubric questions are covered by the participant-submitted questions. Due to budget constraints, TREC assessors could only judge a subset of the question pairs (rubric question, participant question). To reduce the pool of question pairs that needed to be judged, we used two models [32] (`Qwen3-Embedding-8B`[4] and `Qwen3-Reranker-8B`[5]), each of which independently selected the most similar/relevant participant question for each rubric question.

[2]https://trec-rag.github.io/annoucements/2025-rag25-corpus/#-ms-marco-v21-segmented-corpus
[3]https://mediabiasfactcheck.com/
[4]https://huggingface.co/Qwen/Qwen3-Embedding-8B
[5]https://huggingface.co/Qwen/Qwen3-Reranker-8B

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

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

We scored runs using these assessor-assigned labels (ℓ). For topic t, let Q_t be rubric questions, and let $w_q \in \{4, 2, 1\}$ be the importance weight of rubric question $q \in Q_t$. For run r, let $\mathcal{P}_{r,t}$ denote the set of submitted questions. We mapped assessor labels to numeric similarity scores as:

$$g(\ell) = \begin{cases} 1 & \ell = \textsc{very similar}, \\ 0.5 & \ell = \textsc{similar}, \\ 0 & \ell \in \{\textsc{different}, \textsc{very different}\}. \end{cases}$$

Let $\ell_{r,t}(q, p)$ be the judged label for rubric question q and submitted question p. We score run r on topic t by rewarding only the best-matching submitted question per rubric question:

$$S_{r,t} = \frac{1}{W_t} \sum_{q \in Q_t} w_q \max_{p \in \mathcal{P}_{r,t}} g\big(\ell_{r,t}(q, p)\big), \quad W_t = \sum_{q \in Q_t} w_q.$$

Unjudged pairs were assigned VERY DIFFERENT (zero credit). In this setting, the same submitted question could be labeled as similar to multiple rubric questions and rewarded by design, as we wanted to measure the rubric coverage of participant questions.

Regarding the second requirement mentioned in Section 2.2, 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 `gpt-oss-120b` to automatically identify and filter out compound questions (11.3%). 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. With this satisfactory performance, we adopted this automated filtering mechanism to remove compound questions from the submitted question list ($\mathcal{P}_{r,t}$) before scoring. 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 with lower similarity scores.

3.3 Report Assessment

Similar to question assessment, the key idea of report assessment is to check how many rubric answers are covered by each report. These rubric answers function as exemplars of the key information we want a report to convey, rather than templates that systems must match verbatim. For topic t, each rubric question $q \in Q_t$ has a set of short answers $\mathcal{A}_{t,q}$. For a report from run r, the primary assessor assigned a label $\ell_{r,t}(a) \in \{\textsc{supports}, \textsc{partial}, \textsc{contradicts}, \textsc{none}\}$ to each rubric answer a, i.e., using rubric answers as a checklist.

- **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 defined label-to-value mappings for *supportive* and *contradictory* scoring:

$$v_{\text{sup}}(\ell) = \begin{cases} 1 & \ell = \textsc{supp.}, \\ 0.5 & \ell = \textsc{part.}, \\ 0 & \ell \in \{\textsc{contr.}, \textsc{none}\}, \end{cases} \qquad v_{\text{con}}(\ell) = \begin{cases} 1 & \ell = \textsc{contr.}, \\ 0 & \ell \in \{\textsc{supp.}, \textsc{part.}, \textsc{none}\}. \end{cases}$$

We computed topic-level supportive and contradictory scores as:

$$S^{\text{sup}}_{r,t} = \frac{1}{W_t} \sum_{q \in Q_t} \frac{w_q}{|\mathcal{A}_{t,q}|} \sum_{a \in \mathcal{A}_{t,q}} v_{\text{sup}}\big(\ell_{r,t}(a)\big),$$

$$S^{\text{con}}_{r,t} = \frac{1}{W_t} \sum_{q \in Q_t} \frac{w_q}{|\mathcal{A}_{t,q}|} \sum_{a \in \mathcal{A}_{t,q}} v_{\text{con}}\big(\ell_{r,t}(a)\big),$$

where $W_t = \sum_{q \in Q_t} w_q$. Ranking of runs is by the average supportive score over 30 topics (higher is better); contradictory (lower is better) is reported as a secondary diagnostic.

3.4 Per-topic Scores and Headroom

Figure 1 summarizes per-topic rubric-weighted scores for the runs submitted to the TREC 2025 DRAGUN track. For each topic, the dark bar shows the best run, and the light extension shows additional headroom up to a ceiling formed by pooling, for each rubric element, the strongest match across all submissions ("All Runs Combined"). The lines indicate mean scores across runs; for Task 2, we plot both the mean supportive score (red) and the mean contradictory score (orange).

For Task 1 (Question Generation), scores vary substantially across topics, and the best-performing run differs by topic. At the same time, the pooled ceiling is high for most topics and, because each rubric is capped at ten questions, the corresponding "super run" still satisfies the 10-question submission constraint. This suggests that the rubrics largely reflect investigative directions that current systems can generate, but that different systems cover different subsets of these directions. The gap between the best single run and the pooled ceiling, therefore, represents feasible headroom for more robust, article-adaptive question generation.

For Task 2 (Report Generation), supportive coverage is consistently lower than in Task 1, reflecting the added difficulty of retrieving evidence and compressing it into a 250-word, well-attributed

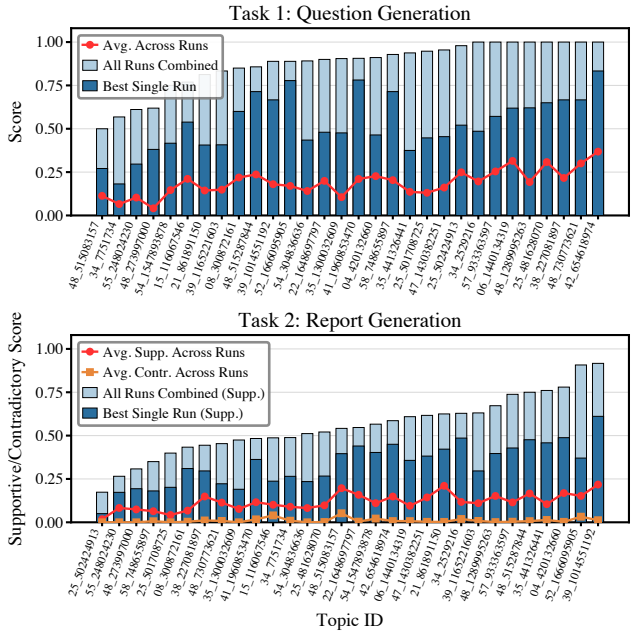

Figure 1: Per-topic rubric-weighted scores. Dark bars: best run for each topic; light: pooled upper bound. Topic IDs omit the `msmarco_v2.1_doc_` prefix.

report. The gap between the best run and the pooled ceiling indicates clear headroom in retrieval, evidence selection, and writing under a strict length budget. Unlike Task 1, this pooled ceiling is generally not attainable by a single report, since combining all supported elements would typically exceed the length limit, so it should be interpreted as an information-availability ceiling across existing systems. Importantly, the average contradictory scores are much smaller than the supportive scores (orange vs. red), suggesting that explicit contradictions to rubric answers are not a major issue for existing runs and do not meaningfully distinguish them. Nevertheless, contradiction labels remain informative: they flag report content that is undesirable to include, even if current differences between runs are driven primarily by what rubric-relevant information systems support rather than what they contradict.

4 AutoJudge: Reusable Automatic Assessment

The DRAGUN rubrics and human judgments provide a high-quality but *static* evaluation of the 2025 submitted runs. To make the collection reusable for evaluating *future* systems on the same tasks, we additionally provide an AutoJudge that can score new runs against the rubrics without requiring new assessor effort. Recent work in IR has shown that LLMs can serve as effective judges for rubric-like assessment and can preserve system rankings derived from humans (e.g., by reporting high rank correlation using Kendall's τ) [21, 23].

We implemented a few-shot AutoJudge using OpenAI's open-weight `gpt-oss-120b`[6] with `temperature` set to 0 and `top_p` set to 1 (other capable LLMs should also work). The prompt mirrors the assessment instructions used for TREC assessors and includes

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

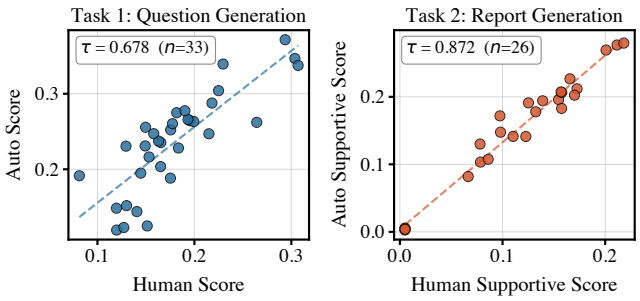

Figure 2: Correlation between human and LLM-based automatic assessments at the run level (averaged score across 30 topics). Each point represents a run.

labeled examples from our organizer baseline runs [30] (4 baselines for Task 1 and 2 baselines for Task 2, excluded from later validation). The model outputs the same labels as the human judging protocol, and we computed run scores using the same scoring scripts as Sections 3.2 and 3.3. On an NVIDIA RTX PRO 6000 GPU, it took roughly 13 hours to assess all 77,880 question pairs and 780 reports. For validation, we report (1) run-level rank correlation via Kendall's τ and (2) label-level agreement between AutoJudge and the human labels using Cohen's κ [4] and Gwet's AC1 [9]. We report AC1 because κ is known to be deflated under heavy class imbalance (the prevalence paradox) [2, 6].

For each rubric question in Task 1 evaluation, we showed the judge the target article, the rubric question, and 40 baseline questions with assessor-provided similarity labels: *very similar*, *similar*, *different*, or *very different*. Baseline questions not selected for human judging were treated as *very different*, matching the track scoring where non-judged pairs received zero credit. The AutoJudge then labeled each (rubric question, participant question) pair, and we computed rubric-weighted run scores after filtering out compound questions. At the run level, the automatic ranking moderately matches the official human ranking (Kendall's $\tau = 0.678$, $n = 33$; Figure 2 left). At the label level, we collapsed *different* and *very different* into a single *no-credit* class (since both yield 0 points) and obtained 82.1% raw agreement, with $\kappa = 0.472$ and AC1 = 0.785.

In terms of report evaluation, for each topic, we provided the judge with the target article, the complete rubric (questions and short answers), and two baseline reports labeled at the rubric-answer level. For a candidate report, the judge assigned one of {*supports, partial, contradicts, none*} for each rubric answer, and we computed the supportive score using the same procedure as Section 3.3 (contradictory scores are similarly low across runs, and therefore not distinguishable). The resulting run-level ranking is highly aligned with the human-derived ranking (Kendall's $\tau = 0.872$, $n = 26$; Figure 2 right). At the answer level, we observed 86.7% raw agreement, with $\kappa = 0.50$ and AC1 = 0.85.

Our rank correlations are in line with recent IR auto-judging studies that compare LLM-derived and human-derived system orderings. For example, LLM-based support judging in the TREC 2024 RAG Track reports run-level Kendall's τ around 0.8 [23], and AutoNuggetizer-style nugget evaluation reports run-level Kendall's τ between 0.727 and 0.901 depending on the manual reference

condition [21]. Overall, our AutoJudge provides the missing component needed to make the released rubrics a reusable benchmark: future systems can be scored consistently with the official human-evaluated leaderboard, enabling direct follow-up experimentation beyond the original track runs.

5 Reusable Resources

To support replication of the DRAGUN track and follow-up research on assistive RAG for news trustworthiness assessment, we release a reusable package of data, judgments, and evaluation code. Most artifacts are available in our public GitHub repository (https://github.com/trec-dragun/resources). The repository README documents the terms of use, citation instructions, file formats, and directory structure for all released artifacts, along with detailed tutorials on how to use the released package. For resources that are conventionally hosted elsewhere (e.g., raw run files on TREC's website[7]), the repository provides pointers and detailed instructions for obtaining them. Table 2 summarizes the collection size and composition. The released package includes the following components:

- Topic set and assessor rubrics.
  - Topics (30 news articles): the set of target articles selected from the MS MARCO V2.1 Document Corpus.
  - Rubrics (30): importance-weighted rubrics created by TREC assessors, one per topic, consisting of questions with one or more expected short answers. Each short answer is supported by one or more reference URLs.
- Human judgments, with assessment guidelines.
  - Task 1 (Question Generation): assessor judgments of question similarity between rubric questions and participant questions.
  - Task 2 (Report Generation): assessor judgments of whether a report supports, partially supports, contradicts, or does not address each rubric answer.
- Participant submissions, with participation guidelines.
  - Runs: all participating teams' submissions.
  - Baseline system: an iterative multi-agent baseline RAG system implementation covering both tasks.
- LLM-based AutoJudge.
  - AutoJudge system: a few-shot prompting-based LLM judge (using `gpt-oss-120b`) that can score additional runs beyond those assessed during the original track.
  - LLM-based assessments: Assessments from the AutoJudge of existing runs for both tasks.
- Scoring scripts: Python scripts that compute scores from either the human judgments or the LLM-based assessments.

6 Discussion

Our release supports several follow-up research directions.

Benchmarking future systems. The most direct use is to evaluate new systems without requiring additional assessor effort. A key design decision was to have assessors build rubrics using open-web lateral reading, rather than restricting rubric construction to the MS MARCO V2.1 Segmented Corpus. This choice helps ensure that rubrics capture comprehensive and unbiased information beyond what submitted systems retrieve. It avoids a common pitfall of "pool-nuggets-then-judge" RAG evaluation, where the evaluation target (nuggets) is constrained by what participating systems happen to surface and fail to spot. The tradeoff is that some rubric answers may be absent from a fixed retrieval corpus, but the headroom analysis suggests substantial missing coverage even for information that should be discoverable (Figure 1), making these rubrics useful for diagnosing retrieval and synthesis gaps.

[7] https://pages.nist.gov/trec-browser/trec34/dragun/runs/

Table 2: Statistics of the DRAGUN collection.

Topics (News Articles) with Rubrics	30
Rubric Questions (Avg / Rubric: 7.9)	236
Rubric Answers (Avg / Rubric: 18.4; Avg / Question: 2.3)	551
Task 1: Question Generation	
Submitted Runs / Teams / Questions	37 / 10 / 11,100
Compound Questions / Not Compound Questions	11.3 / 88.7 (%)
Total Rubric-participant Question Pairs	87,320
Human-assessed Question Pairs	12,733
Very Similar / Similar / Different / Very Different	8.7 / 12.9 / 16.9 / 61.5 (%)
Task 2: Report Generation	
Submitted Runs / Teams / Reports	28 / 8 / 840
Answer-report Pairs (All Human-assessed)	15,428
Supports / Partial / Contradicts / None	5.0 / 7.8 / 0.8 / 86.4 (%)

Comparing evaluation norms. Our rubric-first workflow (expert rubric creation followed by rubric-based scoring) enables direct comparison with nugget-oriented paradigms that derive evaluation units from pooled system outputs, such as AutoNuggetizer [21] and RUBRIC [5]. DRAGUN makes it possible to quantify how rankings change under these different norms and to study whether report-derived nuggets systematically miss expert-identified angles that matter for news trustworthiness assessment.

Advancing automated judging for RAG systems. Because DRAGUN provides both expert labels and an LLM-based AutoJudge, it can serve as a benchmark for developing stronger judges that better match expert decisions, both at the label level and in preserving system rankings. Our current report evaluation is only on rubric-answer coverage (support and contradiction). Future work could extend the framework with complementary dimensions that are important for RAG, such as citation faithfulness (whether cited evidence actually supports the associated claim), readability, etc.

Scaling rubric creation and using rubrics for training. Finally, DRAGUN demonstrates that expert-authored rubrics are a feasible and effective means of specifying the critical information in a news trustworthiness report. By successfully implementing a complete pipeline, from drafting assessment guidelines to training TREC assessors in lateral-reading techniques, we have obtained high-quality rubrics capable of providing nuanced evaluation for RAG systems. This pipeline can be scaled up to generate larger benchmarks and diverse training datasets, providing the necessary supervision signal to align LLM-based assistants with expert-level investigative behaviors. Such scaling would enable models to retrieve, cite, and synthesize evidence addressing expert-identified angles while penalizing contradictory or unsubstantiated claims.

7 Conclusion

We have made the TREC 2025 DRAGUN Track into a reusable resource for the evaluation of assistive RAG systems that help readers

assess news trustworthiness. Our release packages news articles with importance-weighted rubrics, participant submissions, and human judgments for both Task 1 (Question Generation) and Task 2 (Report Generation). To support evaluation beyond the originally judged runs, we have also created and provided an LLM-based AutoJudge that mirrors the rubric-based judging protocol and produces rubric-coverage labels at the same granularity as the human assessments. When validated against the official human judgments, AutoJudge preserves the run-level ordering well, achieving Kendall's $\tau = 0.678$ for Task 1 and $\tau = 0.872$ for Task 2. Together, these resources enable reproducible benchmarking of future systems for lateral-reading-style assistance, and they provide a concrete testbed for advancing automated RAG evaluation using expert rubrics and human labels as a reference point.

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] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–36. doi:10.1257/jep.31.2.211

[2] Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46, 5 (1993), 423–429. doi:10.1016/0895-4356(93)90018-V

[3] Michael Chan, Jingjing Yi, Cristian Vaccari, and Masahiro Yamamoto. 2025. A cross-national examination of the effects of accuracy nudges and content veracity labels on belief in and sharing of misleading news. *Journal of Computer-Mediated Communication* 30, 4 (06 2025), zmaf009. doi:10.1093/jcmc/zmaf009

[4] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104

[5] Naghmeh Farzi and Laura Dietz. 2024. Pencils Down! Automatic Rubric-based Evaluation of Retrieve/Generate Systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval* (Washington DC, USA) *(ICTIR '24)*. Association for Computing Machinery, New York, NY, USA, 175–184. doi:10.1145/3664190.3672511

[6] Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 6 (1990), 543–549. doi:10.1016/0895-4356(90)90158-L

[7] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545. doi:10.1073/pnas.1920498117

[8] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. doi:10.1162/tacl_a_00454

[9] Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48. doi:10.1348/000711006X126600

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[11] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. doi:10.1126/science.aao2998

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[13] Jimmy Lin and Dina Demner-Fushman. 2006. Will Pyramids Built of Nuggets Topple Over?. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, New York City, USA, 383–390. https://aclanthology.org/N06-1049/

[14] Chang Lu, Bo Hu, Qiang Li, Chao Bi, and Xing-Da Ju. 2023. Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *J Med Internet Res* 25 (29 Aug 2023), e49255. doi:10.2196/49255

[15] Gregory Marton and Alexey Radul. 2006. Nuggeteer: Automatic Nugget-Based Evaluation using Descriptions and Judgements. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, New York City, USA, 375–382. https://aclanthology.org/N06-1048/

[16] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. doi:10.18653/v1/2020.acl-main.173

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

[18] Miriam J. Metzger and Andrew J. Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics* 59 (2013), 210–220. doi:10.1016/j.pragma.2013.07.012

[19] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4551–4558. doi:10.24963/ijcai.2021/619

[20] Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review* (June 2020). doi:10.37016/mr-2020-024

[21] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 180–190. doi:10.1145/3726302.3730090

[22] Lisa Scharrer, Marc Stadtler, and Rainer Bromme. 2019. Judging scientific information: Does source evaluation prevent the seductive effect of text easiness? *Learning and Instruction* 63 (2019), 101215. doi:10.1016/j.learninstruc.2019.101215

[23] Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. Assessing Support for the TREC 2024 RAG Track: A Large-Scale Comparative Study of LLM and Human Evaluations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 2759–2763. doi:10.1145/3726302.3730165

[24] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. doi:10.18653/v1/N18-1074

[25] Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. 2023. The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative. *Information Systems Frontiers* 26, 2 (April 2023), 663–688. doi:10.1007/s10796-023-10390-w

[26] Ellen M. Voorhees. 2003. Evaluating Answers to Definition Questions. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*. 109–111. https://aclanthology.org/N03-2037/

[27] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. doi:10.1126/science.aap9559

[28] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 422–426. doi:10.18653/v1/P17-2067

[29] Sam Wineburg and Sarah McGrew. 2019. Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record* 121, 11 (2019), 1–40. doi:10.1177/016146811912101102

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

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

[32] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176 [cs.CL] https://arxiv.org/abs/2506.05176

[33] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 46595–46623. https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf