DS@GT ARC at LongEval: Citation Integrity and Factual Grounding in Scientific QA

Brandon Michaels[1,*], Brendon Johnson[1,*]

[1]Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

Abstract

This paper describes DS@GT ARC's submission to the CLEF 2026 LongEval Task 4 on Retrieval-Augmented Generation (RAG). In this submission, we examine a divergence between traditional natural language evaluation metrics and citation integrity as applied to RAG QA systems. We evaluate a corrective pipeline using Corrective RAG (CRAG) and CiteFix against baseline and frontier model benchmark RAG QA scores. While frontier models maximized answer relevance and fluency scores, our RAGAs LLM-as-judge diagnostics indicate that frontier models would correctly identify relevant documents without using their context in answer generation. Conversely, by filtering chunks pre-generation and enforcing strict entailment of generated claims to the cited material post-generation, our corrective pipeline marginally improved citation faithfulness and answer grounding. We propose that evaluation of trustworthy RAG QA requires metrics that reward strict answer grounding.

Keywords

Retrieval-Augmented Generation, Scientific Question Answering, Large Language Models, Attributed Text Generation

1. Introduction

Scientific knowledge evolves through both continual research and publication of new papers that update previous empirical results, propose novel approaches, or replace old knowledge altogether. While this has led to the discovery of many new research findings, it also has consequential implications for RAG-related question answering tasks. With rapid developments in LLM technology, information can become quickly outdated or even become a complete distraction during question answering or information retrieval tasks [1]. While RAG is an intuitive choice for answering scientific questions since it can base its output on the content of source documents instead of being limited to model memory, it still has ample retrieval related mistakes prompting for stronger, grounded QA. Such mistakes include retrieval of irrelevant documents or paragraphs, missing evidence, wrong document citations, and retrieval of inaccurate or outdated information. These issues become particularly relevant in the domain of science because an answer must be factually grounded, empirically backed by research documents, accurate and currently relevant, and the evidence used must justify the claims.

The 2026 LongEval-RAG [2, 3] task considers this issue in relation to temporally evolving scientific documents from the Connecting Repositories (CORE) scientific literature corpus [4]. For every query, the task organizers present a set of ten documents to consider, and require the participants to output an answer in natural language form, supported by references to the documents used to generate an evidence based answer. Although this inherently simplifies the initial retrieval process by narrowing the candidate set to a smaller region of documents before answer generation, the task organizers included irrelevant documents which still required evidence related filtering. Even within a chunked document, not all information can be considered relevant or highly relevant, which still makes this RAG task meaningful in both natural language answer quality and faithfulness.

As the task supplied a set of pre-retrieved documents for each query, we focused our architecture on post-retrieval optimization of grounded answer generation. To address the challenge of distractor documents, we compare a hybrid chunk retrieval baseline against a strict corrective pipeline. The

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21-24, 2026, Jena, Germany

*Corresponding author.

bmichaels6@gatech.edu (B. Michaels); bjohnson436@gatech.edu (B. Johnson)

0009-0008-2056-6884 (B. Michaels); 0009-0002-1777-2456 (B. Johnson)

corrective architecture utilizes Corrective RAG (CRAG) [5] to filter chunks prior to generation, and CiteFix [6] to enforce strict claim attribution by pruning ungrounded citations post-generation.

Apart from the automated pipelines, we also examine two manually curated runs that use state-of-the-art frontier models (GPT-5.5 Thinking and Claude Opus 4.7 Thinking). While these approaches achieved high scores on standard NLP metrics (ROUGE, BERT) and LLM judged answer relevancy, our RAGAs-based faithfulness evaluation reveals that these same frontier models often fail to properly ground answers in cited materials. The remainder of this paper outlines the related work, details our methodologies, and highlights the gap between generative fluency and citation integrity in our evaluation results.

2. Related Work

Standard RAG frameworks treat source documents as static repositories of knowledge. However, document corpora often exhibit longitudinal context drift, where knowledge mutates over time [7]. This poses a challenge to long-context extraction, where temporal misalignment can contaminate retrieval. Traditional RAG operates under a naive assumption that retrieved contexts are inherently trustworthy, leaving generators vulnerable to retrieval failures. Corrective RAG (CRAG) introduced an active, intermediate evaluation layer that attempts to sanitize the retrieved context before generation [5]. Another challenge for RAG systems is the strict attribution of sources. At generation, an LLM can generate coherent text using supplied context but include ungrounded statements or inaccurate citations [8]. Recent efforts to address this failure mode in trustworthy RAG focus on post-hoc citation mapping to force explicit citation of generated claims [9, 6].

3. Task and Experimental Methodology

3.1. CORE Scientific Document Corpus

The LongEval 2026 Task 4 (LongEval-RAG) dataset provides a textual query alongside 10 pre-retrieved document IDs from the CORE corpus. These 10 retrieved documents are meant to represent the knowledge boundary for generating the answer. This constraint tests a RAG system's ability to filter evidence, isolate distractors, and successfully produce extractive answers. For each document ID, paper title, abstract, and full text fields are made available.

Generative pipelines produced for the LongEval-RAG task must produce outputs aligned with a strict JSONL schema. For each query, systems ingest a JSON containing the query and list of pre-retrieved document IDs. The pipeline processes this input and appends an answer including the generated response and cited documents in the pre-retrieved array. An example of the format is provided in Figure 1.

3.2. Candidate Processing and Retrieval

Given that an initial set of candidate documents is provided by the organizers, we do not aim to create a global corpus retriever. Instead, our architecture focuses on post-retrieval optimization to rank and sanitize the provided payload.

We first segment documents into overlapping semantic chunks. Each chunk is enriched with its parent document's metadata to enable proper attribution. We then execute a second-stage retrieval over these chunks using a hybrid retriever. Our retriever uses a combination of BM25 and BGE (bge-base-en-v1.5) dense vector embeddings to rank the chunks. The top-k chunks based on the input query are fed into the context window of the generator.

3.3. Answer Generation and Post-Hoc Correction

We evaluate three different generative pipelines against the LongEval Organizers' naive baseline.

```
{
  "metadata": {
    "team_id": "LongEval DS@GT",
    "run_id": "example-run",
    "narrative_id": "aa42e210a...",
    "narrative": "How can a device avoid futile access..."
  },
  "references": [275699672, 275699915, 122371639, ...],
  "answer": [
    {
      "text": "A device can avoid futile NR access by using...",
      "citations": [0, 1]
    }
  ]
}
```

Figure 1: Expected JSONL schema for LongEval-RAG submissions. Note that the `citations` array relies on local positional indices mapping back to the `references` array, rather than the global document IDs in the generated text.

3.3.1. DS@GT Hybrid Baseline

This architecture represents a standard RAG implementation. The top-k chunks retrieved by the BM25+BGE hybrid ranker are injected directly into the prompt for the Gemma-4-31B instruction-tuned model. The model synthesizes an initial answer and outputs an array of cited document indices.

3.3.2. Corrective RAG and CiteFix

Our primary research pipeline introduces active interventions to the Hybrid Baseline.

1. Pre-Generation (CRAG): Before generation, a lightweight cross-encoder (cross-encoder/ms-marco-MiniLM-L-6-v2) evaluates retrieved chunks. Chunks below the entailment threshold are dropped from the supplied context.
2. Post-Generation (CiteFix): After generation, we use CiteFix to parse generated claims against cited document chunks. Citations lacking entailment to support the generated claims are pruned. Figure 2 shows the steps of the CiteFix pipeline.

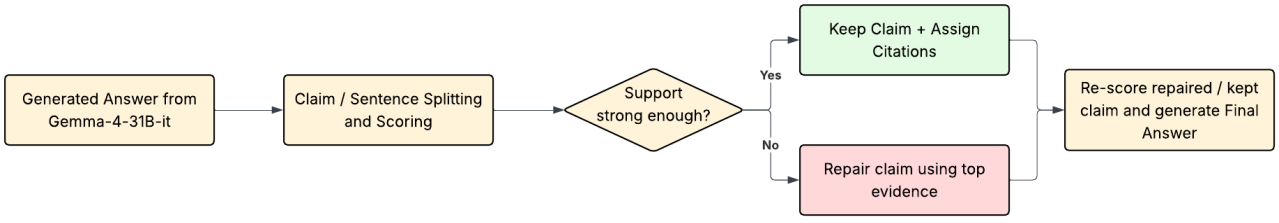

Figure 2: CiteFix Evidence Validation Repair Pipeline

3.3.3. Frontier Model Benchmarks

To establish an upper bound on long-context parametric reasoning, we bypass scoring and chunking entirely. We prompt ChatGPT-5.5 Thinking and Claude Opus 4.7 Thinking to synthesize answers directly from the raw text of the 10 candidate documents.

3.4. Reference-Free RAGAs Evaluation

In the absence of a gold QA set with ground truth answers in the development phase, we utilized an LLM-as-judge setup (Claude Sonnet 4.6) via the RAGAs framework [10]. The evaluation calculated the faithfulness and relevancy metrics using multi-step prompting algorithms with the judge LLM. The scoring workflows are presented in Figures 3 and 4.

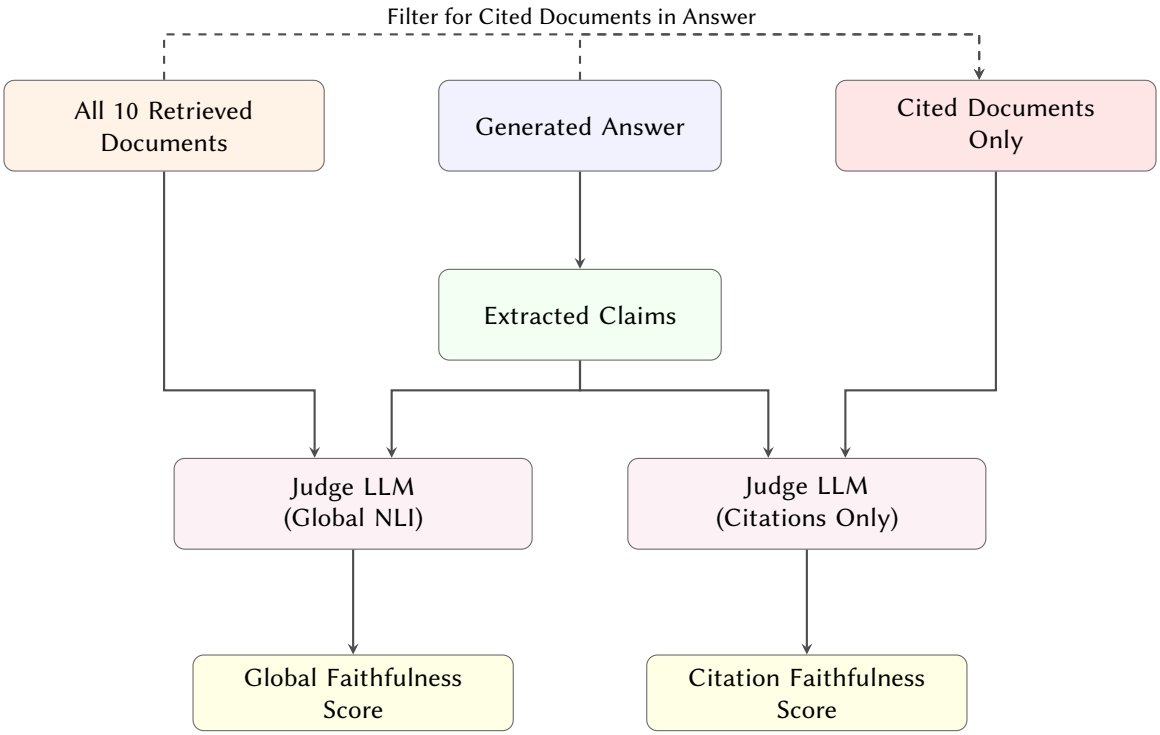

Figure 3: RAGAs Faithfulness workflows. Global Faithfulness verifies claims extracted from the generated answer against the entire 10-document context array to penalize ungrounded statements. Citation Faithfulness filters for only cited documents, forcing the judge to verify claims strictly against the documents cited in the answer.

Global Faithfulness This metric measures the factual consistency of the generated response against the 10 pre-retrieved documents. Using RAGAs, the generated answer is decomposed into discrete, standalone claims. The judge LLM (Claude Sonnet 4.6) then decides if claims in the generated answer are grounded within the entirety of the supplied documents. The final metric is the ratio of grounded statements to total statements in answers on the evaluation set.

Citation Faithfulness While Global Faithfulness evaluates the entire task context, Citation Faithfulness isolates the documents claimed to have been used at generation. During this evaluation pass, answer claims are evaluated against only cited documents. Large gaps between Global Faithfulness and Citation Faithfulness reveal a failure mode where grounded answers are produced but falsify or otherwise hallucinate the citation.

Answer Relevancy This metric quantifies the completeness of the answer with respect to the query. Rather than use semantic similarity of question and answer, this metric uses the judge model to produce counterfactual queries based on only the generated answer. The similarity of the generated queries and the original query is then used to score each answer. This metric rewards query-aligned answers, but systematically penalizes safe abstention when the generator receives no context or context not relevant to the original query.

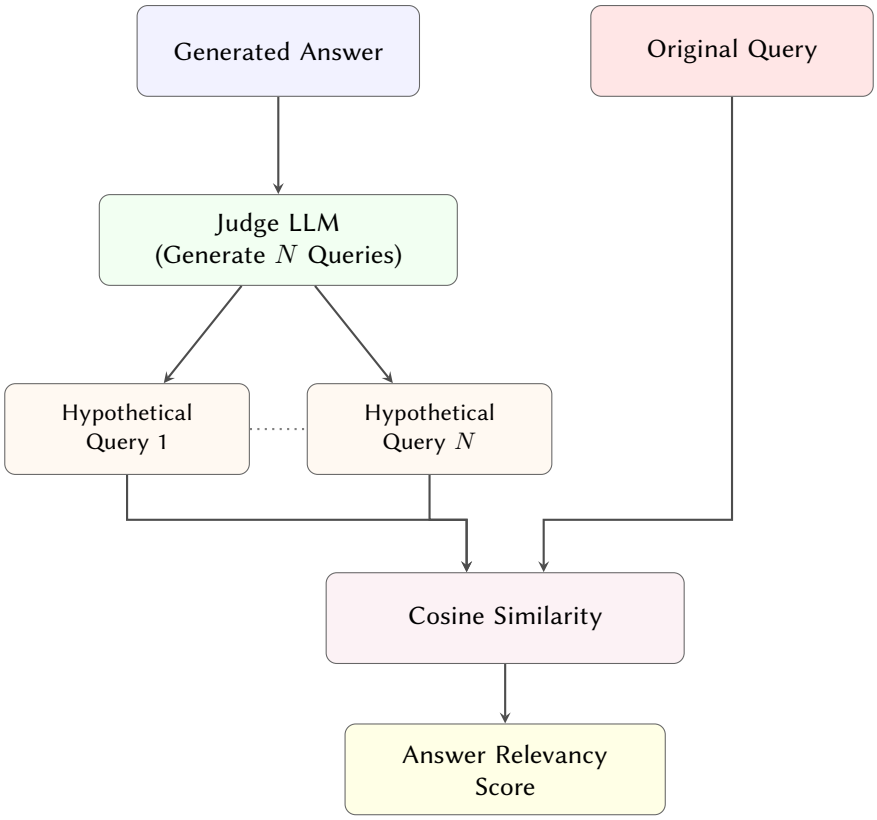

Figure 4: RAGAs Answer Relevancy workflow. Only the generated answer is provided to the judge model in order to generate counterfactual queries. These are compared to the original user query via embedding cosine similarity.

3.5. Official Task Evaluation Metrics

While we performed our own suite of evaluation tests, the LongEval Organizers also evaluated submissions against a hidden gold set of answers. The headline metrics evaluate output quality using ROUGE and BERT scores. ROUGE measures structural lexical overlap by calculating the longest common subsequence (LCS) and n-gram intersection between submitted and gold answers. Conversely, BERT Score evaluates semantic equivalence by calculating cosine similarity of token embeddings. Higher metric scores indicate the generated answer shares higher lexical or semantic similarity with the gold answer.

Additionally, the organizers calculate citation precision, which evaluates whether the documents cited by our pipeline match those selected in the gold set. Let C_q be the set of documents cited by our pipeline for a given user query q, and let R_q be the set of documents labeled relevant in the gold standard. Citation precision calculates the ratio of correctly cited relevant documents to the total number of cited documents ($|C_q \cap R_q|/|C_q|$). The final score is averaged over all queries. The organizers provide detailed metrics covering natural language coherency, consistency, and clarity, as well as a variety of other retrieval heuristics measuring the prompt in comparison to the input query. The full list of the results including the additional IR axiom metrics can be viewed in the Appendix.

4. Results

Here, we report the official LongEval-RAG evaluation metrics and our own reference-free RAGAs diagnostics for each of our pipelines.

4.1. Official Task Metrics

Tables 1 and 2 detail aggregate performance of the pipelines on the official task metrics. The frontier model runs establish a high upper bound for natural language generation. GPT-5.5 achieved the highest similarity scores with a ROUGE$_{F1}$ of 0.188 and a BERT$_{F1}$ of 0.227. Claude Opus 4.7 achieved a perfect Citation Precision (1.000) and the highest Nugget Coverage (0.443), the best performance in selecting the correct document IDs from each pre-retrieved set of documents.

We see that our hybrid chunk reranking benchmark significantly outperformed the Organizer's Naive Baseline across all lexical and semantic metrics. Our corrective intervention pipeline using CRAG+CiteFix experienced a drop in both ROUGE and BERT scores compared to the hybrid baseline. This is an artifact of the abstention behavior which suppresses generation of ungrounded text.

System	$\mathbf{ROUGE_{F1}}$	$\mathbf{BERT_{F1}}$	Precision	Nugget Cov.	TFC1
(LongEval Lab) Naive Baseline	0.082	-0.118	0.200	0.124	-0.758
(DS@GT) Hybrid Baseline	0.132	0.118	0.766	0.210	-0.322
CRAG+CiteFix	0.126	0.076	0.692	0.152	-0.505
GPT-5.5 Thinking	**0.188**	**0.227**	0.904	0.308	**0.434**
Claude Opus 4.7 Thinking	0.169	0.157	**1.000**	**0.443**	0.017

Table 1
LongEval Task 4 results across automatic and manually curated runs. Manual systems provide an approximate upper bound for natural language metrics.

Table 3 shows system performance on secondary metrics evaluating sentence-level consistency and clarity based on TREC-AutoJudge Axioms [11, 12]. Claude Opus 4.7 produced the most coherent output (COH1 $= 0.857$). The Naive Baseline scored well on CLAR2, indicating high readability despite poor accuracy.

System	$\mathbf{ROUGE_1}$	$\mathbf{ROUGE_2}$	$\mathbf{ROUGE_L}$	$\mathbf{BERT_P}$	$\mathbf{BERT_R}$
(LongEval Lab) Naive Baseline	0.142	0.023	0.082	-0.313	0.090
(DS@GT) Hybrid Baseline	0.207	0.023	0.132	0.117	0.117
CRAG+CiteFix	0.201	0.024	0.126	0.064	0.086
GPT-5.5 Thinking	**0.292**	**0.062**	**0.188**	**0.242**	0.210
Claude Opus 4.7 Thinking	0.284	0.057	0.169	0.097	**0.217**

Table 2
Detailed gold-answer similarity metrics. R-1 and R-2 are Rouge Scores for lexical similarity between generated and gold set on word-one-grams and bi-grams respectively. R-L is Rouge Score for longest overlapping phrase. BERT-P and BERT-R, are BERT precision and recall respectively for all queries. GPT-5.5 manual outputs achieve the strongest $ROUGE_1$, $ROUGE_2$, $ROUGE_L$, and $BERT_P$ scores, while Claude Opus 4.7 Thinking manual outputs achieve the highest $BERT_R$ score.

System	COH1	COH2	CLAR1	CLAR2
(LongEval Lab) Naive Baseline	-0.150	-0.064	-0.093	0.678
(DS@GT) Hybrid Baseline	0.024	0.020	-0.006	-0.030
CRAG+CiteFix	-0.021	**0.021**	0.013	-0.183
GPT-5.5 Thinking	0.427	0.004	**0.014**	0.239
Claude Opus 4.7 Thinking	**0.857**	-0.119	0.013	**0.818**

Table 3
Claude Opus 4.7 Thinking manual responses are strongly preferred on sentence-level coherency and clarity (COH1 and CLAR2). GPT-5.5 Thinking and CRAG+CiteFix approach lead on secondary metrics (CLAR1 and COH2 respectively). The organizer naive baseline is weak on coherence but scores highly on CLAR2.

System	Global Faithfulness	Citation Faithfulness	Answer Relevancy
Hybrid Baseline	0.741	0.741	0.531
CRAG+CiteFix	**0.784**	**0.758**	0.476
GPT-5.5 Thinking	0.559	0.417	**0.756**
Claude Opus 4.7 Thinking	0.553	0.528	0.646

Table 4
RAGAs automated evaluation metrics across generation architectures. CRAG+CiteFix maximizes factual grounding and attribution accuracy, while frontier models (Opus, GPT-5.5) prioritize answer relevancy at the expense of strict citation integrity.

4.2. RAGAs Evaluation Metrics

Table 4 shows the results of our reference-free RAGAs evaluation. These metrics evaluate strictly for attribution integrity and grounding of claims rather than lexical overlaps or semantic similarity. Our CRAG+CiteFix pipeline maximized faithfulness to the supplied context, achieving a Global Faithfulness of 0.784 and a Citation Faithfulness of 0.758. Conversely, the frontier models performed quite poorly on these evaluations. Faithfulness scores for both GPT-5.5 Thinking and Claude Opus 4.7 Thinking collapsed despite producing answers that were highly relevant to the queries. This indicates a failure mode where the models output highly relevant answers without extracting their claims from the cited documents.

5. Discussion

We observe a tradeoff between several performance metrics and our RAGAs-based faithfulness and citation accuracy evaluation when comparing structured corrective pipelines with unconstrained frontier models. While the frontier models (GPT-5.5, Opus 4.7) achieved higher official leaderboard scores, our RAGAs diagnostics (Table 4) indicate a failure to ground the answers in the retrieved context. Faced with noisy context, the frontier models generated answers predominantly from pre-trained knowledge.

For example, GPT-5.5 generally led the leaderboard scores and had the highest Answer Relevancy of 0.756, but exhibited extremely low Citation Faithfulness at 0.417 indicating low reliance on cited contexts in answers. From the leaderboard precision of 0.904 and manual trace audits, we find that while GPT-5.5 is highly capable of identifying the relevant document(s) within a noisy context window, the details of generated answers typically did not rely on that same context.

The corrective pipeline achieved Global Faithfulness of 0.784 and Citation Faithfulness of 0.758, which were leading on our submissions. However, this triggered penalties on some of the leaderboard metrics. When retrieved chunks lacked the answer, our corrective pipeline would abstain to answer rather than generate an answer from parametric memory. These abstentions are penalized on metrics reliant on relevancy or n-gram overlaps despite exhibiting a more faithful failure mode than parametric memory leaks. This highlights a challenge in trustworthy RAG benchmarking where chosen evaluation metrics may reward failure modes which hallucinate citations to maintain conversational fluency over safer abstentions.

To evaluate the statistical significance of the CRAG + CiteFix pipeline, we conducted a Wilcoxon signed-rank test on the paired RAGAs evaluation scores across the 47 test queries. The results indicated that the corrective interventions did not yield a statistically significant improvement to citation faithfulness or answer relevancy.

6. Future Work

A few limitations to our evaluation should be noted. While our frontier model submissions score well on lexical similarity metrics, they score poorly on faithfully grounding answers in the retrieved evidence.

Conversely, our Gemma-4-31B pipelines exhibit relatively high faithfulness but much lower relevance and citation precision. As our prompting strategy did not require answers to be strictly extractive, instead requesting synthesis across relevant passages, we posit that forcing output text to be extractive would greatly improve lexical similarity and potentially induce more faithful answering from the higher capacity models, if at a penalty to clarity and coherence.

An immediate next step for this work is to verify that the CRAG and CiteFix corrective interventions successfully improve answer grounding across families and sizes of generators. Our initial results do not indicate a statistically significant increase in faithfulness over the baseline. Further exploration of NLI model use for scientific claim verification is needed to understand if the null result is driven by NLI brittleness rather than applicability of post-hoc pruning in this domain. Additionally, the lack of a CRAG+CiteFix submission to the leaderboard using GPT-5.5 Thinking or Claude Opus 4.7 Thinking was due to practical submission limit and time constraints within the evaluation period. We encourage release of the hidden gold set query answers to facilitate expanded ablation studies.

Our RAGAs evaluations are not without limitations. A single, uncalibrated reference-free LLM judge was used to evaluate answer faithfulness and relevancy. Future evaluations will deploy multi-judge ensembles, enabling the evaluation of intraclass correlations across judges and mitigation of known evaluator biases such as self-preference or verbosity.

Finally, the described RAG systems and RAGAs evaluators should be extended across multiple snapshots within the LongEval corpus. The current evaluation focused on a static payload. Applying these pipelines across a longitudinal dataset would assess QA performance under the types of temporal drift which penalize use of point-in-time parametric memory. Additionally, the threshold hyperparameters for CRAG and CiteFix are tuned for a specific generation task. The performance of these techniques over multiple snapshots may degrade, similar to standard retrievers and generators. Longitudinal evaluation of these systems should explore adaptive tuning of the active interventions, as well as assess how corpus shifts impact the faithfulness and relevancy judgments of the underlying RAGAs LLM judge. As our faithfulness and relevancy metrics are tied to an LLM-as-a-judge method, temporal bias should be evaluated and compared to other known judge biases when interpreting longitudinal RAGAs scores.

7. Conclusions

We find a divergence between traditional language metrics and factual grounding in RAG. While unconstrained frontier models generally achieved better ROUGE and BERT score with query-aligned answers, our diagnostics show this comes from utilization of parametric memory rather than retrieved context. Conversely, we find that deployment of active interventions pre- and post-generation do improve structural honesty in RAG QA systems. Ultimately, securing RAG QA pipelines in scientific domains rely on both strict grounding to evidence and overall quality of answer. We therefore recommend adoption of evaluation metrics that reward structural safety and proper failure modes within RAG systems.

Acknowledgments

We thank the Data Science at Georgia Tech Applied Research Competitions (DS@GT ARC) group for their support. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA [13].

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3.1 Pro in order to aid in formatting, grammar, and spell checking. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Advances in Neural Information Processing Systems, 2020.

[2] S. Bouaraba, T. Breuer, M. Cancellieri, A. El-Ebshihy, M. Fröbe, P. Galuscáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, P. Mulhem, F. Piroi, D. Pride, P. Schaer, D. Schwab, Overview of the clef 2026 longeval lab on longitudinal evaluation of model performance, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science (LNCS), Springer, 2026.

[3] S. Bouaraba, T. Breuer, M. Cancellieri, A. El-Ebshihy, M. Fröbe, P. Galuscáková, L. Goeuriot, G. Iturra-Bocaz, J. Keller, P. Knoth, A. K. Kruff, P. Mulhem, F. Piroi, D. Pride, P. Schaer, D. Schwab, Extended overview of the clef 2026 longeval lab on longitudinal evaluation of model performance, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[4] P. Knoth, D. Herrmannova, M. Cancellieri, L. Anastasiou, N. Pontika, S. Pearce, B. Gyawali, D. Pride, Core: A global aggregation service for open access papers, Nature Scientific Data 10 (2023) 366.

[5] S.-Q. Yan, J.-C. Gu, Y. Zhu, Z.-H. Ling, Corrective retrieval augmented generation, arXiv preprint arXiv:2401.15884 (2024). URL: https://arxiv.org/abs/2401.15884.

[6] H. Maheshwari, et al., Citefix: Enhancing RAG accuracy through post-processing citation correction, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Industry Track, 2025. URL: https://aclanthology.org/2025.acl-industry.23/.

[7] J. Keller, T. Breuer, P. Schaer, Evaluation of temporal change in ir test collections, in: Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 3–13. URL: https://doi.org/10.1145/3664190.3672530. doi:10.1145/3664190.3672530.

[8] N. Liu, T. Zhang, P. Liang, Evaluating verifiability in generative search engines, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 7001–7025. URL: https://aclanthology.org/2023.findings-emnlp.467/. doi:10.18653/v1/2023.findings-emnlp.467.

[9] T. Gao, H. Yen, J. Yu, D. Chen, Enabling large language models to generate text with citations, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6465–6488. URL: https://aclanthology.org/2023.emnlp-main.398/. doi:10.18653/v1/2023.emnlp-main.398.

[10] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/. doi:10.18653/v1/2024.eacl-demo.16.

[11] J. H. Merker, M. Fröbe, B. Stein, M. Potthast, M. Hagen, Axioms for retrieval-augmented generation, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, ACM, Padua, Italy, 2025, pp. 67–77. URL: https://doi.org/10.1145/3731120.3744601. doi:10.1145/3731120.3744601.

[12] N. Farzi, T. Hagen, E. Yang, M. Fröbe, R. Pradeep, H. A. Rahmani, X. Wang, O. Zendel, M. P. L. Dietz, Auto-judge: A cross-task benchmark for comparing llm judges for citation-grounded rag systems, 2026.

[13] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: http://www.pace.gatech.edu.

8. Appendix

8.1. Answer Generation Prompt

For each query, the LLM receives both the user scientific query and the CRAG-filtered evidence block for the research based approaches (excluded for baseline). The prompt instructs the model to produce a concise, paraphrased answer that synthesizes evidence from the retrieved document chunks without directly copying. The exact prompt template used is below:

```
Question:
{query}

CRAG-filtered evidence:
{evidence_block}

Write a direct answer in 1-3 sentences, about 30-65 words total.

Rules:
- No bullets, markdown, or inline citations.
- Do not mention passage numbers or evidence labels.
- Synthesize the answer rather than copying source text.
- Even if evidence is partial, give the strongest concise answer
  supported by the evidence.
- Write in English only.
- Use standard English spellings for chemical names and technical terms.
```

8.2. TREC-AutoJudge IR Axiom Results

Metric	Naive Baseline	Hybrid Baseline	CRAG+CiteFix	GPT-5.5 Thinking	Claude Opus 4.7 Thinking
GEN-TFC1	-0.758	-0.322	-0.505	**0.434**	0.017
GEN-LNC1	0.000	0.001	**0.005**	0.000	0.000
GEN-REG	-0.184	-0.126	-0.257	**0.345**	0.065
GEN-AND	**0.737**	-0.263	-0.242	-0.157	-0.136
GEN-DIV	**0.626**	0.223	0.406	-0.512	-0.290
GEN-STMC1	-0.662	-0.112	-0.364	**0.501**	-0.124
GEN-STMC2	0.041	0.018	-0.008	0.018	**0.042**
GEN-PROX1	**0.137**	0.006	0.002	0.014	-0.007
GEN-PROX2	-0.083	-0.008	-0.004	-0.029	**0.055**
GEN-PROX3	**0.116**	0.000	0.001	0.001	-0.015
GEN-PROX4	**0.136**	0.000	0.000	0.002	0.013
GEN-PROX5	**0.050**	0.000	0.006	0.025	0.019
GEN-ASL	0.000	0.000	0.000	0.000	0.000
GEN-TF-LNC	-0.005	-0.010	-0.013	**0.009**	-0.008

Table 5
Results for TREC-AutoJudge IR axiom metrics. The GEN-* metrics evaluate proximity properties of generated answers with respect to the input query. Higher values indicate better results.

Tables 5 and 6 detail our submission results on TREC-AutoJudge IR Axioms. Notably, the organizers' naive baseline performed best across several proximity and intersection-based metrics. We hypothesize that the performance gap is an artifact of the divergence between extractive and abstractive generation methods. Classical IR axioms like GEN-AND and GEN-PROX4 reward lexical overlap and term proximity. An extractive baseline will inherently satisfy these conditions. Our RAG pipelines were prompted to perform abstractive synthesis across the supplied context. This process can break lexical overlaps and alter term distances, resulting in penalty by these axioms. We acknowledge that stricter extractive prompting could improve scores on these axioms.

Metric	Naive Baseline	Hybrid Baseline	CRAG+CiteFix	GPT-5.5 Thinking	Claude Opus 4.7 Thinking
COH1-0.75	-0.150	0.024	-0.021	0.427	**0.857**
COH2-0.75	-0.064	0.020	**0.021**	0.004	-0.119
COV1-SE-0.5	**0.859**	-0.150	-0.047	-0.404	0.165
COV2-SE-0.5	**0.363**	-0.024	0.005	-0.051	-0.107
COV3-SE-0.5	**0.018**	-0.017	-0.003	-0.025	-0.032
CONS3-0.5	**0.005**	0.003	**0.005**	0.003	**0.005**
CORR1-0.75	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CLAR1-0.5	-0.093	-0.006	**0.013**	**0.013**	**0.013**
CLAR2-0.5	0.678	-0.030	-0.183	0.239	**0.818**

Table 6
Appendix results for TREC-AutoJudge coherence, coverage, consistency, correction, and clarity metrics.