Adapting AutoARGUE for Automatic Report Evaluation under Missing Citation Annotations

Divrose Kaur **Jatin Bedi** **Jasmeet Singh**
Thapar Institute of Engineering & Technology
divrose19@gmail.com jatin.bedi@thapar.edu jasmeet.singh@thapar.edu

Abstract

We adapt the AutoARGUE framework (Walden et al., 2026) for Task A.2 of RAG4Reports 2026, which requires ranking 57 report generation systems across 68 topics using automated evaluation. The RAGTIME-1 corpus poses a fundamental challenge: all nugget annotations use a `no-reference-doc` sentinel rather than ground-truth document citations, rendering the original citation-relevance gating inoperable. We address this with three adaptations: automatic sentinel detection with forced direct LLM-based nugget matching; a WEAK POSITIVE partial credit mechanism for sentences that correctly answer nuggets but lack attesting citations; and a report-level request alignment check. Our `nugget_coverage_weighted` metric achieves the highest topic-level Pearson correlation (r=0.599) of any non-coordinator submission, closely approaching the coordinator baseline (r=0.607).

1 Introduction

Automatic evaluation of retrieval-augmented generation (RAG) systems is an active area of research, with nugget-based frameworks emerging as the leading paradigm (Walden et al., 2026; Mayfield et al., 2024). RAG4Reports 2026 Task A.2 operationalises this paradigm at scale: given 57 machine-generated reports across 68 topics from the RAGTIME-1 corpus, participants must produce a system ranking that correlates with human annotation rankings, measured by Kendall's τ (Fagin et al., 2004).

The task organizers designated AutoARGUE (Walden et al., 2026) as the official baseline. AutoARGUE implements the ARGUE decision tree, which evaluates each report sentence by checking whether its citations attest its claims and whether it answers key nugget questions. Our work extends this baseline to handle the RAGTIME-1 corpus, which differs fundamentally from the NeuCLIR corpus (Lawrie et al., 2025) that AutoARGUE was designed for: every nugget annotation uses the sentinel `no-reference-doc` instead of a ground-truth document ID, making citation-relevance lookup inoperable.

We describe three principal adaptations, our evaluation setup, and results on the official leaderboard.

2 Task and Data

Task definition. Task A.2 receives as input 57 JSONL files of machine-generated reports (one per system, 68 topics each) from TREC RAGTIME 2025 participants, human-curated nugget files (QA pairs, one per topic), and a `report-requests.jsonl` containing user backgrounds and information needs. The output is a TSV of per-topic metric scores; the primary evaluation metric is gap-aware rank correlation (Fagin et al., 2004) on the macro-averaged F1 rankings against human annotation.

Corpus. The RAGTIME-1 retrieval corpus comprises four multilingual JSONL files: English news articles (`eng-docs`), and Russian-, Chinese-, and Arabic-to-English translations (`rus-trans`, `zho-trans`, `arb-trans`).

3 Methodology

3.1 The `no-reference-doc` Problem

The ARGUE decision tree (Mayfield et al., 2024) at Node B asks: *Is the cited document relevant?* This is answered by looking up the cited document ID in a doc-to-nugget annotation map — each nugget answer is linked to specific documents that attest it. In RAGTIME-1, every nugget answer uses `no-reference-doc` as its citation, intentionally omitting ground-truth annotations. This causes Node B to always return *No*, collapsing every cited sentence to a negative outcome and producing zero

Proceedings of the First Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports), pages 103–107

scores for all systems.

Our solution: (1) detect the all-sentinel condition automatically at load time; (2) strip `no-reference-doc` from all nugget citation lists before building the doc-to-nugget map; (3) enable `always_check_all_nuggets` mode, forcing the LLM to evaluate every nugget against every sentence regardless of citation annotation. This bypasses annotation-based gating and substitutes direct LLM judgment throughout, at the cost of increased LLM call volume per topic.

3.2 Adapted Decision Tree

Figure 1 illustrates our adapted ARGUE decision tree. Table 1 summarises the outcome cases; the principal changes from the original framework are the WEAK POSITIVE branch and the report-level alignment check.

Condition	Outcome	Credit
Cited + attested + nugget match	Strong Positive	1.0
Cited + attested + no nugget	Neutral	0.0
Cited + not attested + nugget	Weak Positive	0.5
Cited + not attested + no nugget	Negative	0.0
No citation + requires citation	Negative	0.0
No citation + no citation needed	Neutral	0.0
Report misaligned with request	Align. mult.	$\times 0.7$

Table 1: Adapted decision tree outcome matrix.

WEAK POSITIVE partial credit. During development, we observed that some systems correctly answered nuggets — demonstrating genuine information coverage — but received F1 = 0.0 because their cited documents were not attested by the corpus. This discards a meaningful quality signal: the system found relevant information but cited imperfectly. We implement $0.5\times$ partial credit in sentence precision for WEAK POSITIVE outcomes, making the scorer sensitive to information coverage independent of citation quality.

Request alignment check. The ARGUE framework states that reports should be tailored to the specific user's information need. We add a report-level alignment check: the full report text is evaluated against the user background and problem statement from `report-requests.jsonl` by the LLM judge. Misaligned reports receive a $0.7\times$ multiplier on their F1 score. In practice, most systems scored alignment $= 1.0$; the multiplier correctly flagged near-empty or off-topic reports.

3.3 Scoring Metrics

$$\textit{sent\_sup} = \frac{\sum_s \text{credit}(s)}{|\text{sentences requiring citation}|} \tag{1}$$

$$\textit{nug\_cov} = \frac{|\text{nuggets answered}|}{|\text{total nuggets}|} \tag{2}$$

$$\text{F1} = \frac{2 \cdot \textit{sent\_sup} \cdot \textit{nug\_cov}}{\textit{sent\_sup} + \textit{nug\_cov}} \times \textit{alignment} \tag{3}$$

The primary submission metric is `f1_macro`: the mean F1 across all evaluated topics per system.

3.4 LLM Judge and Evaluation Setup

We use Llama-3.3-70B Instruct (Grattafiori et al., 2024) (4-bit quantisation, `unsloth/Llama-3.3-70B-Instruct-bnb-4bit`) served locally via vLLM (Kwon et al., 2023), matching the judge model used in the AutoARGUE paper (Walden et al., 2026). Prior to local vLLM deployment, we evaluated several provider options. Cloud-based APIs (Llama-3.3-70B and smaller variants via Groq) were constrained by rate limits that made full-dataset evaluation infeasible. Locally-hosted smaller models via Ollama (`qwen2.5:7b`) lacked the nuanced judgment capability required for reliable nugget matching. These constraints motivated local deployment of the full 70B model. Document text is truncated to 800 characters for LLM attestation calls to reduce per-topic compute cost.

Due to resource constraints, we evaluated 5 representative topics per system (1001: AI/technology, 1009: science/weather, 1017: entertainment, 1069: technology history, 1083: environmental science) rather than all 68.

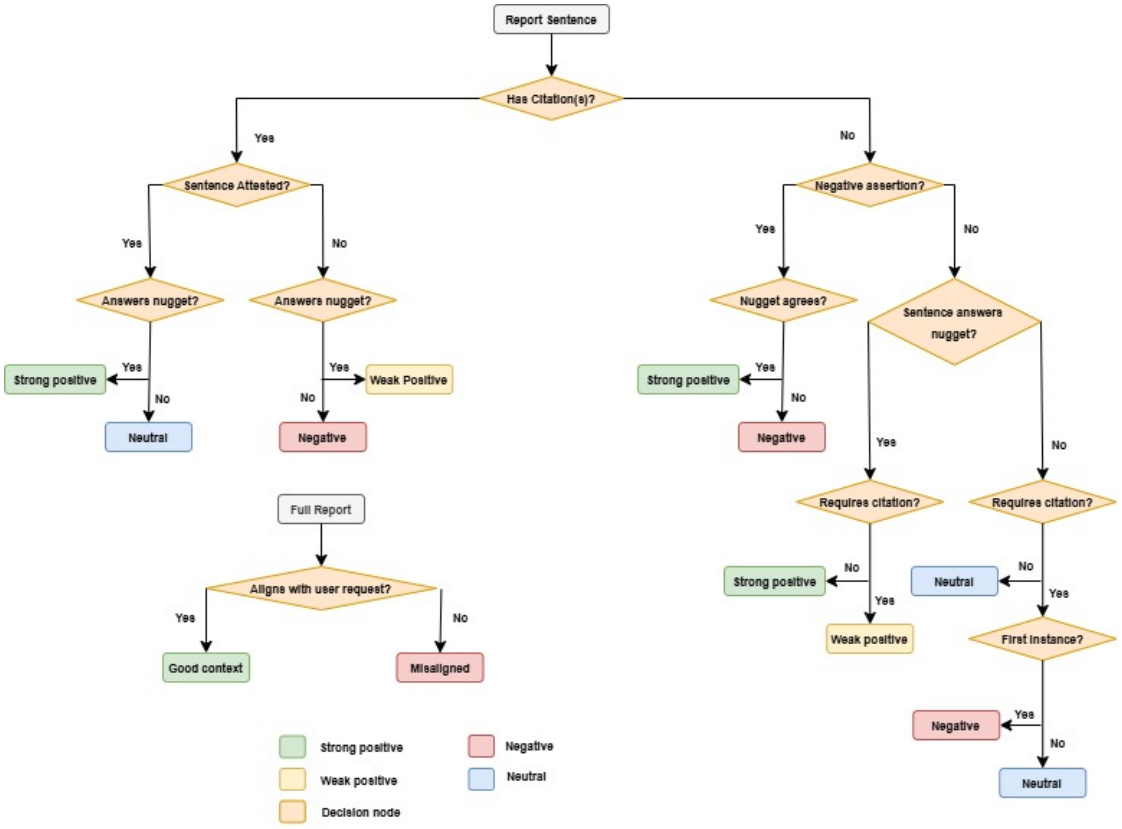

Figure 1: Adapted ARGUE decision tree for RAGTIME-1. The left subtree handles sentences *with* citations; the right subtree handles sentences *without* citations. The bottom section shows the report-level request alignment check, which applies a $0.7\times$ score multiplier to misaligned reports.

4 Results

4.1 Leaderboard Results

Table 2 shows our submissions against the coordinator baseline on the Task A leaderboard.

Team	Metric	tau_gap	τ	r
Coord.	autoargue-f1	0.470	0.636	0.607
Ours	f1_weighted	0.127	0.334	0.573
Ours	nugget_cov_w	0.075	0.289	**0.599**
Ours	sentence_support	−0.056	0.183	0.247

Table 2: Task A results. tau_gap is the primary metric (higher is better); r = topic-level Pearson. `nugget_cov_w` achieves the highest topic-level Pearson of any non-coordinator submission.

Our `nugget_coverage_weighted` metric achieves the highest topic-Pearson correlation (r=0.599) of any non-coordinator submission, closely approaching the coordinator baseline of 0.607. `sentence_support` alone performs below random (tau_gap < 0), confirming that citation quality in isolation is insufficient to rank systems without the nugget coverage signal.

4.2 System-Level Score Distribution

Our pipeline produced clear differentiation across the 57 systems, with F1 macro ranging from 0.611 (`friend-proceed`) to 0.000 for systems with zero citations or zero nugget coverage. Notably, `possibility-prime` had 27 citations but 0.000 nugget coverage — citing documents without addressing the information need. Citation support was consistently lower than sentence support across all systems, indicating a pervasive pattern of citing documents that do not fully attest their sentences — a known failure mode in RAG systems.

4.3 Per-Topic Analysis

Table 3 shows per-topic scores for the top system (`friend-proceed`). Topic 1069 (LiveJournal) was consistently the hardest across all evaluated systems, suggesting ambiguous or poorly-specified requests. Topic 1083 (mammoth extinction) achieved 100% nugget coverage for this system.

Topic	F1	Nug.	Sent.	Cit.
1001 (AI)	0.683	0.600	0.792	0.522
1009 (Weather)	0.750	0.667	0.857	0.289
1017 (Disney)	0.587	0.500	0.711	0.378
1069 (LiveJrn.)	0.296	0.364	0.250	0.222
1083 (Mammoth)	0.741	1.000	0.588	0.250
Aggr.	**0.611**	0.626	0.640	0.332

Table 3: Per-topic scores for top system (`friend-proceed`). Nug. = nugget coverage; Sent. = sentence support; Cit. = citation support.

5 Analysis

Nugget coverage vs. F1. The higher topic-Pearson of `nugget_coverage_weighted` (r=0.599) compared to `f1_weighted` (r=0.573) suggests nugget recall is a stronger per-topic signal than the combined F1. That `sentence_support` alone yields a negative tau_gap (-0.056) confirms this component is unreliable in isolation, while nugget coverage directly measures information coverage.

Partial evaluation scope. The strong score differentiation across systems (F1 macro 0.00–0.61) suggests the 5-topic sample was sufficient to surface the most salient quality differences. Full 68-topic evaluation was infeasible under our resource constraints.

The `no-reference-doc` adaptation. Our sentinel detection and direct LLM-matching approach generalises to any dataset with incomplete annotation coverage. The primary tradeoff is increased LLM call volume, as every nugget must be checked against every sentence regardless of document overlap.

Sentence support as a standalone signal. The negative tau_gap for `sentence_support` (-0.056) confirms that citation quality alone is an unreliable ranking signal. The F1 harmonic mean combining nugget coverage and sentence support is substantially more reliable than either component alone.

6 Conclusion

We adapted AutoARGUE for the RAGTIME-1 corpus by addressing the `no-reference-doc` sentinel problem, adding WEAK POSITIVE partial credit, and implementing report-level request alignment checking. Our `nugget_coverage_weighted` metric achieves the highest topic-Pearson of any non-coordinator submission (r=0.599), closely approaching the coordinator baseline. These adaptations generalise to evaluation scenarios with incomplete nugget annotation coverage.

Limitations

Evaluation covers 5 of 68 topics per system due to resource constraints; full coverage may alter system rankings at the margins. Document truncation to 800 characters is effective for inverted-pyramid news text but may not generalise to corpora with different information density. The request alignment multiplier ($0.7\times$) was set heuristically; a calibrated value may perform better. The 4-bit quantised Llama judge may produce different judgments than the full-precision model used in the official coordinator evaluation.

Ethical Considerations

This work develops automated evaluation metrics for RAG systems. Automated metrics influence system development and benchmarking; over-reliance on any single metric without human validation carries the risk of rewarding systems that optimise for the metric rather than genuine quality. Our results show that individual metric components (e.g., sentence support alone) can be poor proxies for human judgments, underscoring the importance of multi-faceted evaluation.

Acknowledgments

This work was conducted as part of undergraduate research at Thapar Institute of Engineering & Technology. The authors thank the RAG4Reports 2026 organizers for their support and guidance throughout the shared task.

References

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. 2004. Comparing and aggregating rankings with ties. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 47–58.

Grattafiori et al. 2024. The Llama 3 herd of models. *arXiv preprint*.

Woosuk Kwon, Zhuowei Li, Siyuan Zhu, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS*

29th Symposium on Operating Systems Principles, pages 611–626.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2025. Overview of the TREC 2024 NeuCLIR track. In *Proceedings of the Thirty-Third Text REtrieval Conference (TREC 2024)*.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1904–1915.

William Walden, Marc Mason, Orion Weller, Laura Dietz, John Conroy, Neil Molino, Hannah Recknor, Bryan Li, Gabrielle Kaili-May Liu, Yu Hou, Dawn Lawrie, James Mayfield, and Eugene Yang. 2026. Auto-ARGUE: LLM-based report generation evaluation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*.