Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 520-533
DOI: https://doi.org/10.51903/jtie.v4i2.535

Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations

Jiaying Jin*[1]
Email: binghua.zhou@yahoo.com
[1]*Applied Analytics, Columbia University, NY, USA*
*Corresponding Author

Abstract
Retrieval-augmented generation (RAG) reduces unsupported generation by grounding answers in source content, but retrieval alone does not guarantee that every output claim is attributable to evidence. This paper presents Evidence-Chain Reliable RAG, an empirical hallucination-detection and provenance framework that scores whether generated response sentences are supported by the corresponding RAG source record. The evaluation uses the complete RAGTruth JSONL data available for this study: 2,965 source records, 17,790 assistant responses, and 14,289 exact-offset hallucination spans across Data2Text, QA, and summarization. The experiment converts word-level spans into response-level, sentence-level, and character-span targets; extracts lexical, BM25, TF-IDF, unsupported-number, unsupported-entity, refusal, and Evidence-Chain Score features; and compares seven methods. On the official held-out test split of 2,700 responses, RandomForest achieved the best case-level F1 of 0.626 and PR-AUC of 0.553. The proposed ECS-Span model achieved case-level F1 of 0.614, ROC-AUC of 0.742, and PR-AUC of 0.536 while also producing deterministic provenance explanations. At sentence level, RandomForest again achieved the highest F1 of 0.321; the proposed method obtained F1 of 0.312, ROC-AUC of 0.777, and PR-AUC of 0.245. Exact character-span localization remained difficult, with character-level F1 of 0.197 because sentence-level predictions often include supported text around shorter hallucinated spans. The findings indicate that evidence-chain features are useful for interpretable RAG auditing, but precise span extraction requires token-level sequence labeling or a comparable fine-grained model.

Keywords: *retrieval-augmented generation; hallucination detection; RAGTruth; evidence attribution; trustworthy AI.*

I. INTRODUCTION

Retrieval-augmented generation has become a practical architecture for question answering, summarization, and data-grounded generation because it lets a language model condition on external documents rather than rely only on parametric memory (Lewis et al., 2020). A RAG system first retrieves passages or structured records and then asks a generator to synthesize an answer. This architecture improves coverage and updateability, yet it also creates a reliability problem: a generated response can sound coherent while contradicting the retrieved evidence, adding unsupported facts, refusing when evidence is available, or mixing a correct answer with an unsupported detail.

For user-facing NLP systems, this failure mode is not merely stylistic. It is a source-attribution defect. A RAG answer is reliable only when users can inspect which source evidence supports each output claim. A response-level label can indicate that some problem exists, but it cannot tell a reviewer which sentence or span should be edited. Fine-grained supervision is therefore necessary for both detection and practical remediation. The RAGTruth benchmark was designed for this setting by providing hallucination annotations for RAG outputs at both case and word

Received on Juni 2025; Revised on July 2025; Accepted on August 2025; Published on August 2025

levels (Wu et al., 2024). This paper uses the complete RAGTruth JSONL data available for the experiment and evaluates hallucination detection at three granularities: response, sentence, and character span. The evaluation covers Data2Text, QA, and summarization, which makes it possible to test whether evidence-chain features behave consistently across structured and passage-based sources.

The phrase evidence chain refers to a measurable path from the source record, through the nearest supporting source chunk, to each generated sentence and predicted hallucination span. The proposed framework uses this path in two ways. First, it computes support features such as maximum evidence overlap, local TF-IDF similarity, BM25 support, unsupported numeric details, unsupported named entities, refusal indicators, and a combined Evidence-Chain Score. Second, it returns the nearest evidence chunk and a deterministic natural-language explanation for every flagged sentence.

The contribution is threefold. First, the paper gives a reproducible pipeline for turning RAGTruth span annotations into response-level, sentence-level, and character-level evaluation targets. Second, it compares unsupervised evidence matching, rule-based support-gap scoring, and supervised attribution classifiers under the same train-validation-test protocol. Third, it integrates deterministic provenance explanations with detection results, so that a risk score is accompanied by the closest evidence and a specific support gap. The claims are intentionally cautious: the proposed model is not presented as an overall predictive winner, but as an interpretable reliability layer that can be audited and improved.

II. LITERATURE REVIEW

RAG builds on open-domain retrieval and reading systems such as DrQA, REALM, dense passage retrieval, and late-interaction neural retrieval (Chen et al., 2017; Guu et al., 2020; Karpukhin et al., 2020; Khattab & Zaharia, 2020). Classical lexical retrieval remains relevant because BM25 is efficient and interpretable for passage matching (Robertson & Zaragoza, 2009). Neural retrieval improves semantic matching, but retrieval alone does not ensure that generation follows the retrieved evidence.

Hallucination evaluation in summarization and generation has moved from document-level adequacy toward factual consistency. Maynez et al. (2020) showed that abstractive summaries can be fluent while containing unsupported content. FactCC, FRANK, SummaC, TRUE, and related resources formalized factual consistency as an evaluation target (Honovich et al., 2022; Kryscinski et al., 2020; Laban et al., 2022; Pagnoni et al., 2021). A recurring lesson is that surface similarity is insufficient because a generated sentence can reuse source words while changing a number, entity, relation, or scope.

Natural language inference is often used as a factuality proxy because it predicts entailment, contradiction, and neutrality between a premise and a hypothesis (Bowman et al., 2015; Williams et al., 2018). Transformer encoders such as BERT, RoBERTa, and DeBERTa have provided the backbone for many NLI and hallucination detectors (Devlin et al., 2019; He et al., 2021; Liu et al., 2019). Sentence-BERT and related embedding methods make semantic similarity more efficient, but similarity is still a support signal rather than proof of entailment (Reimers & Gurevych, 2019).

Fact-checking and long-form factuality work provide complementary ideas. FEVER frames factual verification as evidence retrieval plus claim verification (Thorne et al., 2018). RARR revises language-model outputs using retrieval and attribution, while FActScore decomposes long-form output into atomic facts (Gao et al., 2023; Min et al., 2023). SelfCheckGPT estimates hallucination risk from consistency among sampled generations (Manakul et al., 2023). RAGAS evaluates RAG with answer faithfulness and context relevance (Es et al., 2023). These approaches show that reliable generation evaluation requires claim-level units, evidence links, and interpretable diagnostics.

Interpretability research also motivates provenance explanations. LIME argued that users need understandable local explanations rather than only global accuracy (Ribeiro et al., 2016), and Doshi-Velez and Kim (2017) emphasized that interpretability should be evaluated relative to the decision problem. In RAG, the relevant decision is whether the retrieved evidence supports a generated claim. A provenance explanation is useful when it states the flagged claim, shows the nearest evidence, and names the support gap. The evidence-chain framework operationalizes this idea with deterministic templates rather than an evaluated external LLM.

III. RESEARCH METHOD

The evaluation used the complete RAGTruth JSONL data available for this study. The source file contains source records for Summary, QA, and Data2Text tasks. The response file contains assistant responses generated by six model families, official train or test split labels, generator metadata, and a list of hallucination spans when annotations are present. Each span includes start and end character offsets, span text, label type, and metadata explaining the annotation.

Table 1. RAGTruth data audit and integrity checks

item	count
source records	2965
assistant responses	17790
annotated hallucination spans	14289
assistant responses with at least one span	7664
assistant responses without spans	10126
exact span offset matches	14289

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 520-533
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Before modeling, the script verified dataset integrity. The data contain 2,965 source records, 17,790 assistant responses, and 14,289 annotated hallucination spans. All 14,289 span offsets exactly matched the corresponding response substrings. Table 1 reports the data audit, while Table 2 reports the response-level split used for training, validation, and final testing. The official test split was retained for final evaluation. The official train split was divided by source record, not by individual response, into training and validation partitions with random seed 42; this avoids placing responses to the same source in both the training and validation partitions.

Table 2. Response-level split by task and hallucination label

role	task_type	no-span	span	total
Train	Data2Text	1309	2927	4236
Train	QA	2759	1267	4026
Train	Summary	2646	1164	3810
validation	Data2Text	314	748	1062
validation	QA	711	297	1008
validation	Summary	630	318	948
Test	Data2Text	321	579	900
Test	QA	740	160	900
Test	Summary	696	204	900

The response was the prediction unit for case-level evaluation. A response was labeled positive when its span list was nonempty. Sentence-level examples were created by deterministic punctuation and newline segmentation of assistant responses. A sentence was labeled hallucinated when its character range overlapped at least one gold hallucination span. This produced 144,325 sentence instances: 99,119 for training, 24,754 for validation, and 20,452 for held-out testing. Table 3 shows the span-type distribution, and Table 4 shows response labels by generator model.

Table 3. Gold hallucination span label distribution

Span label	Count
Evident Baseless Info	6237
Evident Conflict	5324
Subtle Baseless Info	2527
Subtle Conflict	201

Figure 1 summarizes the pipeline. Each source record was split into evidence chunks. Summary records were segmented primarily by sentence; QA records were split by passage markers; Data2Text records were flattened into key-value and review chunks. For each generated sentence, the system selected the nearest evidence chunk and extracted support and risk features from that pair.

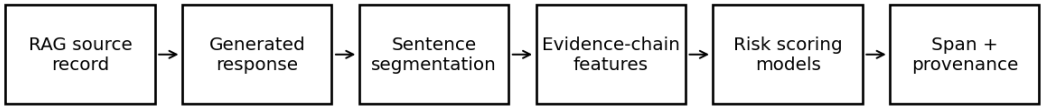

Figure 1. Evidence-chain reliable RAG pipeline used in the experiment

The feature extractor computed maximum Jaccard overlap, response-token recall, source-token precision, normalized BM25 support, local TF-IDF cosine similarity, unsupported-number rate, unsupported-entity rate, refusal indicators, sentence length, source-chunk count, task indicators, and an Evidence-Chain Score. The Evidence-Chain Score combines lexical recall, Jaccard support, TF-IDF support, BM25 support, numeric support, and entity support into a single support estimate. Lower evidence-chain support and higher unsupported-detail rates indicate higher hallucination risk.

Table 4. Response label distribution by generator model

model	no-span	span	total
gpt-3.5-turbo-0613	2564	401	2965
gpt-4-0613	2559	406	2965
llama-2-13b-chat	1288	1677	2965
llama-2-70b-chat	1570	1395	2965
llama-2-7b-chat	1133	1832	2965
mistral-7B-instruct	1012	1953	2965

Seven methods were evaluated, as summarized in Table 5. Lexical-overlap, BM25-evidence, and TF-IDF-cosine are unsupervised support baselines. NLI-proxy is a deterministic support-gap formula that approximates contradiction risk using semantic support, token recall, unsupported numbers, unsupported entities, and refusal cues. LogReg-attribution is a balanced logistic-regression classifier over evidence-chain features. RandomForest is a 60-tree class-balanced random forest. Proposed-ECS-Span is a 40-tree ExtraTrees span classifier using the Evidence-Chain Score, support-gap features, unsupported-detail features, and task indicators.

Table 5. Experimental methods and operational definitions

Method	Type	Operational definition
Lexical-overlap	Unsupervised	Risk = 1 - maximum Jaccard overlap between a response sentence and any evidence chunk.
BM25-evidence	Unsupervised	Risk = 1 - normalized BM25 support score over the source chunks.
TF-IDF-cosine	Unsupervised	Risk = 1 - maximum local TF-IDF cosine similarity.
NLI-proxy	Rule-based	Weighted support-gap proxy using semantic support, token recall, unsupported numbers, unsupported entities, and refusal cues.
LogReg-attribution	Supervised	Balanced logistic regression over evidence-chain attribution features.
RandomForest	Supervised	60-tree random forest with class-balanced bootstrap sampling.
Proposed-ECS-Span	Proposed	40-tree ExtraTrees span classifier using Evidence-Chain Score, support-gap, unsupported-detail, and task features.

Thresholds for all methods were selected on the validation split by maximizing F1 and then frozen for held-out testing. Response-level scores were obtained by taking the maximum sentence risk score within a response. Sentence-level scores were evaluated directly against the hallucination-overlap labels. Character-span localization projected proposed sentence predictions back to response character ranges and compared those ranges with exact gold hallucination spans. This

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 520-533
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

character metric is stricter than case-level or sentence-level detection because it penalizes supported words that occur near an unsupported phrase. The provenance module generates deterministic natural-language explanations. It is template-based: no external LLM is called, and the explanation is not evaluated as a generated LLM output. Each explanation reports the flagged sentence, the nearest evidence chunk, and the main support gap, such as weak evidence-chain score, unsupported number, unsupported entity, or refusal-like wording.

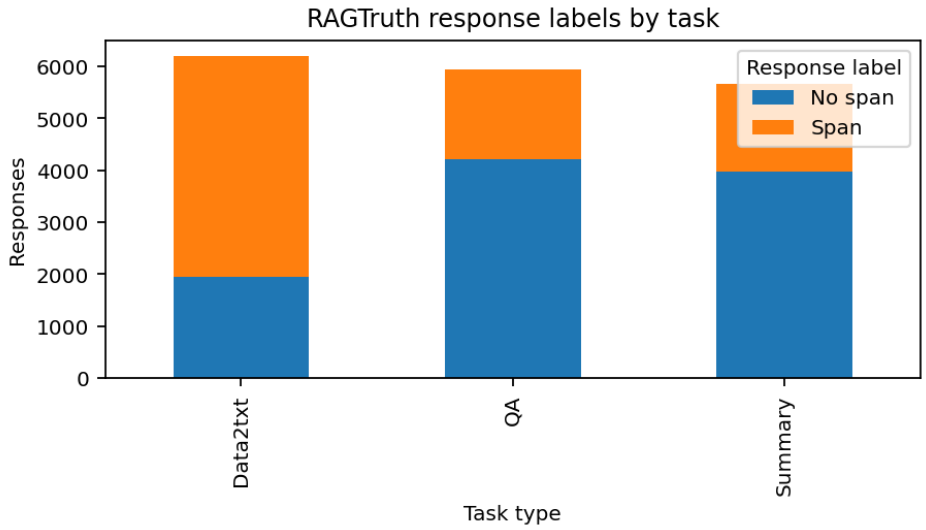

Figure 2. Task distribution and response-level hallucination labels

Figures 2 and 3 visualize the dataset composition. Data2Text has the highest positive response rate, while QA and Summary contain more responses without annotated spans. Evident Baseless Info and Evident Conflict dominate the span labels, showing that many errors introduce unsupported content or directly conflict with the source.

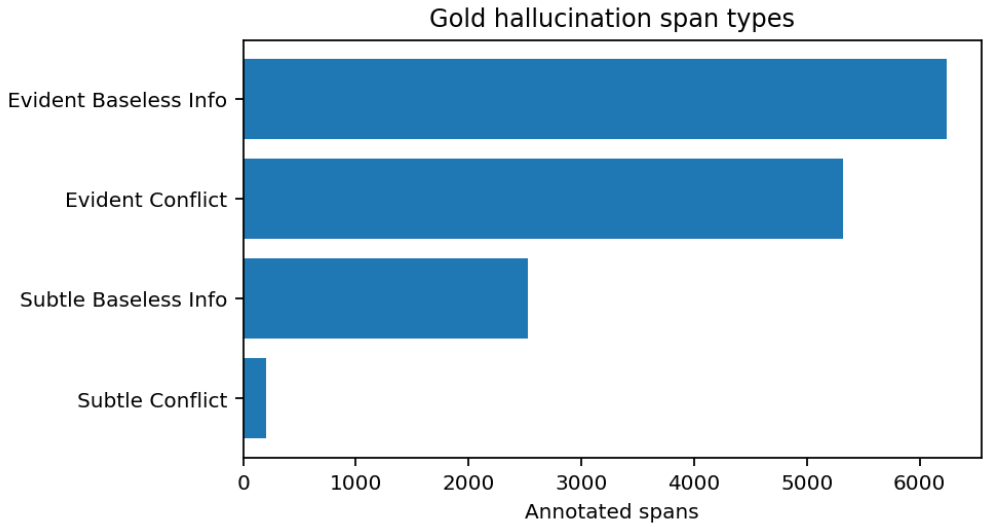

Figure 3. Distribution of annotated hallucination span types

IV. FINDINGS AND DUSCUSSION

A. Result

The complete RAGTruth test set is challenging at fine granularity because it contains all three task types and a lower positive rate at sentence level. In the held-out test split, 943 of 2,700 responses have at least one hallucination span. At sentence level, 1,672 of 20,452 test sentences overlap a gold span. This imbalance explains why response-level recall can look stronger than precise localization.

Table 6. Held-out case-level comparison on response labels

method	accuracy	precision	recall	f1	mcc	roc_auc	pr_auc	ms/item
Lexical-overlap	0.578	0.446	0.862	0.588	0.293	0.671	0.450	0.0001
BM25-evidence	0.543	0.424	0.867	0.570	0.249	0.643	0.435	0.0001
TF-IDF-cosine	0.575	0.446	0.887	0.593	0.305	0.671	0.450	0.0001
NLI-proxy	0.554	0.435	0.928	0.593	0.308	0.682	0.451	0.0001
LogReg-attribution	0.611	0.469	0.870	0.610	0.341	0.711	0.499	0.001
RandomForest	0.643	0.494	0.856	0.626	0.376	0.754	0.553	0.018
Proposed-ECS-Span	0.612	0.471	0.881	0.614	0.350	0.742	0.536	0.010

Case-level results are reported in Table 6. RandomForest achieved the best held-out case-level F1 of 0.626, the best MCC of 0.376, the best ROC-AUC of 0.754, and the best PR-AUC of 0.553. Proposed-ECS-Span achieved case-level F1 of 0.614, ROC-AUC of 0.742, and PR-AUC of 0.536. The proposed method is therefore competitive but not the strongest purely predictive model. Its main contribution is that it returns reusable provenance information together with the score.

Table 7. Held-out sentence-level comparison on hallucination-overlap labels

method	accuracy	precision	recall	f1	mcc	roc_auc	pr_auc	ms/item
Lexical-overlap	0.549	0.120	0.710	0.205	0.134	0.642	0.111	0.0001
BM25-evidence	0.631	0.126	0.592	0.208	0.127	0.641	0.113	0.0001
TF-IDF-cosine	0.626	0.137	0.678	0.228	0.167	0.687	0.127	0.0001
NLI-proxy	0.618	0.143	0.736	0.239	0.190	0.707	0.138	0.0001
LogReg-attribution	0.819	0.207	0.430	0.279	0.207	0.742	0.192	0.0001
RandomForest	0.824	0.234	0.511	0.321	0.259	0.785	0.255	0.002
Proposed-ECS-Span	0.817	0.225	0.505	0.312	0.248	0.777	0.245	0.001

Sentence-level results are reported in Table 7. RandomForest again performed best on F1 and PR-AUC, reaching sentence-level F1 of 0.321 and PR-AUC of 0.255. Proposed-ECS-Span reached F1 of 0.312, ROC-AUC of 0.777, and PR-AUC of 0.245. Logistic regression achieved higher accuracy than unsupervised baselines but lower F1 than the tree-based supervised models. The ROC and precision-recall curves in Figures 4 and 5 show that supervised evidence-chain models separate positive and negative sentences better than pure lexical baselines, although precision remains low because exact positive sentences are relatively rare.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 520-533
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

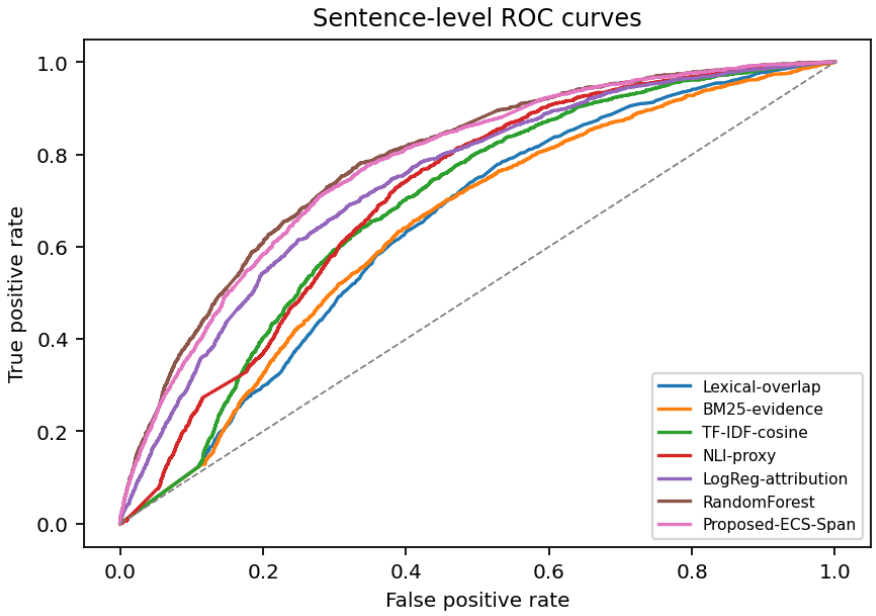

Figure 4. Sentence-level ROC curves for all compared methods

Character-span localization is the most difficult evaluation. Table 8 reports character-level precision of 0.120, recall of 0.559, and F1 of 0.197 for Proposed-ECS-Span under sentence-range projection. The result should be interpreted cautiously. The classifier can often identify a sentence that contains an unsupported phrase, but the projected range includes supported tokens surrounding that phrase. Figure 6 illustrates this overmarking behavior. This is acceptable as an early-stage diagnostic layer, but it is not sufficient for precise editing without a token-level sequence labeling stage.

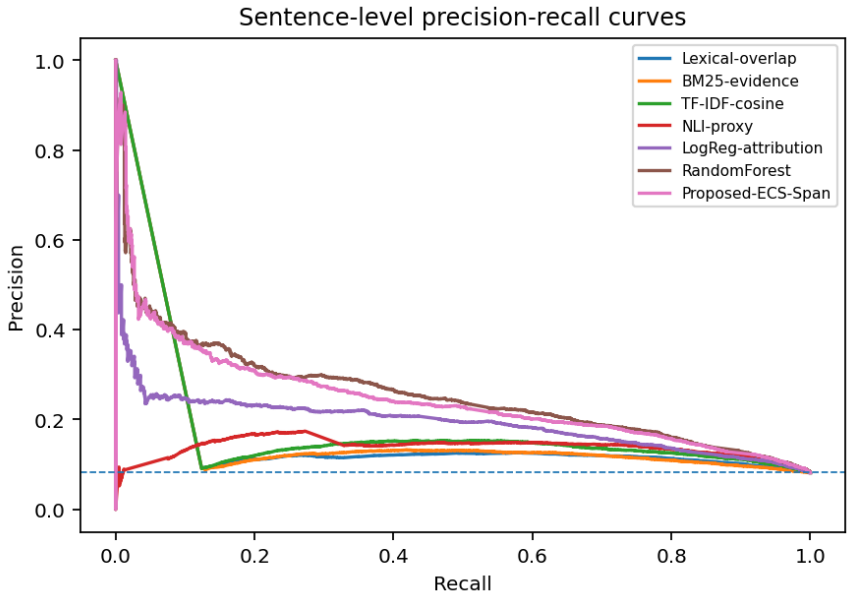

Figure 5. Sentence-level precision-recall curves for all compared methods

The stratified results in Table 9 show that performance varies by task and generator family. The proposed method performed better on QA than Summary by F1, and Data2Text remained

challenging because structured source records often contain many details that are easy to paraphrase or omit. Model-level strata should not be read as a ranking of foundation models; they are used here to diagnose detector behavior under different output distributions.

Table 8. Character-span localization results for the proposed method

method	projection	char P	char R	char F1	mean msg F1
Proposed-ECS-Span	sentence	0.120	0.559	0.197	0.474

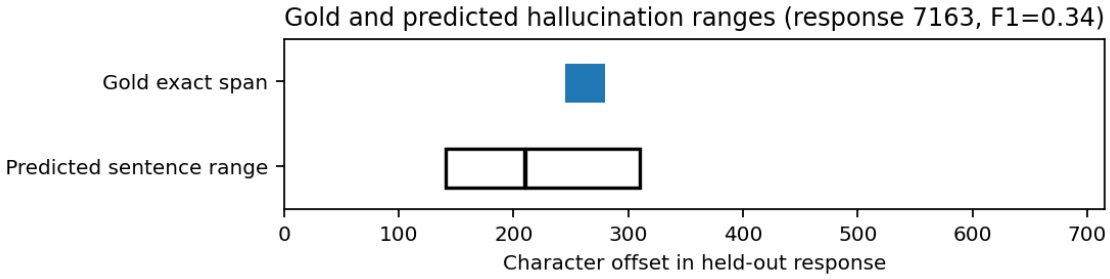

Figure 6. Gold exact hallucination span and predicted sentence range for a held-out response

The provenance module achieved full attribution coverage for flagged test sentences, as reported in Table 10. Coverage means that every flagged sentence was linked to a nonempty nearest evidence chunk. It does not mean that the evidence supports the sentence; instead, the evidence chunk gives a reviewer a concrete source candidate to inspect. Table 11 provides examples of deterministic provenance explanations, and Figure 8 shows the same evidence-chain idea visually. Figure 7 reports local scoring latency, indicating that the methods are lightweight enough for post-generation review.

Table 9. Stratified proposed-method sentence results on held-out data

Stratum	accuracy	precision	recall	f1	roc_auc	pr_auc
Proposed-ECS-Span by task_type=Data2Text	0.699	0.210	0.597	0.311	0.727	0.225
Proposed-ECS-Span by task_type=QA	0.865	0.254	0.531	0.344	0.813	0.302
Proposed-ECS-Span by task_type=Summary	0.942	0.343	0.136	0.195	0.750	0.173
Proposed-ECS-Span by model=gpt-3.5-turbo-0613	0.854	0.062	0.523	0.111	0.795	0.076
Proposed-ECS-Span by model=gpt-4-0613	0.837	0.058	0.562	0.105	0.760	0.051
Proposed-ECS-Span by model=llama-2-13b-chat	0.812	0.283	0.559	0.376	0.793	0.302
Proposed-ECS-Span by model=llama-2-70b-chat	0.794	0.238	0.527	0.327	0.786	0.270
Proposed-ECS-Span by model=llama-2-7b-chat	0.812	0.307	0.487	0.376	0.747	0.296
Proposed-ECS-Span by model=mistral-7B-instruct	0.794	0.325	0.448	0.376	0.750	0.334

B. Discussion

The main empirical finding is that reliable RAG evaluation must separate three questions: whether a response contains any hallucination, which sentence contains it, and what evidence relationship explains the flag. Case-level detection is useful for triage, but it can hide weak localization.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 520-533
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Sentence-level detection narrows the review target, but a sentence can contain both supported and unsupported content. Character-level scoring exposes this limitation and shows why a sentence classifier should not be mistaken for a precise span editor.

Table 10. Provenance attribution coverage for flagged test sentences

Flagged test sentences	Flagged with evidence	Coverage rate	Mean ECS flagged
3753	3753	1	0.247

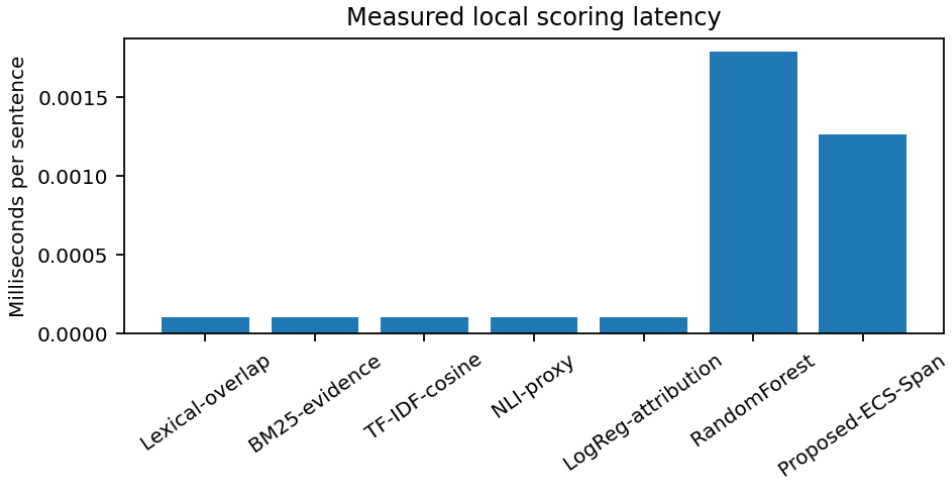

Figure 7. Measured local scoring latency by method

The results also clarify the role of the proposed method. RandomForest is the best predictive model in this experiment for both response-level and sentence-level F1. Proposed-ECS-Span is not claimed to be superior on every metric. Its value is that it combines competitive scoring with an explicit evidence chain, nearest evidence chunk, and deterministic explanation. This makes it better suited to human review workflows than a score-only detector, even when a different classifier has higher F1.

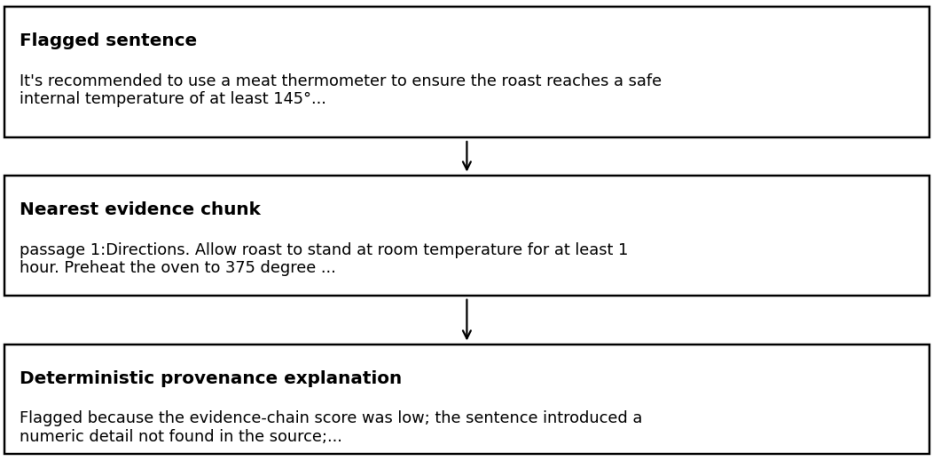

Figure 8. Evidence-chain attribution example with nearest source chunk

The low precision in sentence and character metrics is informative. Many test responses contain long evidence-related sentences, and the gold span may cover only a short unsupported clause. A detector that marks the sentence can still be useful for alerting, but it will overmark supported text. For editing tools, the next step should be a token-level or phrase-level sequence labeler trained to identify exact unsupported spans within sentences already flagged by the evidence-chain model.

Table 11. Example deterministic provenance explanations for held-out predictions

Gold	Pred	Prob	Sentence excerpt	Nearest evidence excerpt	Deterministic provenance explanation
1	1	0.906	It's recommended to use a meat thermometer to ensure the roast reaches a safe internal temperature of at least 145°F for medium-rare, 1...	passage 1:Directions. Allow roast to stand at room temperature for at least 1 hour. Preheat the oven to 375 degree F. Rub roast with House Season...	Flagged because the evidence-chain score was low; the sentence introduced a numeric detail not found in the source; the sentence introduced an en...
1	1	0.905	However, it is recommended to use a meat thermometer to ensure the pork chops reach an internal temperature of at least 145°F (63°C) fo...	passage 3:1 Rinse pork chops, and sprinkle the wet chops on both sides with the spice mixture. With your hands, massage the spice rub into the me...	Flagged because the evidence-chain score was low; the sentence introduced a numeric detail not found in the source; the sentence introduced an en...
1	1	0.892	It should reach an internal temperature of 130°F to 135°F for medium-rare, 140°F to 145°F for medium, and 150°F to 155°F for medium-wel...	passage 1:Directions. Allow roast to stand at room temperature for at least 1 hour. Preheat the oven to 375 degree F. Rub roast with House Season...	Flagged because the evidence-chain score was low; the sentence introduced a numeric detail not found in the source.

C. Limitations and Future Work

Several limitations remain. First, RAGTruth is a fixed benchmark generated from a defined set of model families and source tasks. Results may not transfer directly to newer models, longer contexts, multimodal RAG, or domain-specific retrieval systems. Second, the feature extractor uses deterministic lexical, BM25, TF-IDF, numeric, entity, and refusal cues. These features are reproducible and interpretable, but they do not fully capture discourse-level entailment, causal relations, or subtle contradictions. Third, the provenance explanations are deterministic templates. This avoids overstating the role of LLMs and makes the outputs reproducible, but it also limits explanation richness. Future work could compare deterministic explanations with LLM-generated explanations under a separate human evaluation protocol. Fourth, character-level localization remains weak because sentence projection overmarks supported words. Future work should add token-level sequence labeling, span boundary calibration, and task-specific span projection for Data2Text, QA, and summarization. Finally, validation thresholds were selected by F1.

Deployment settings may require different calibration. A legal, medical, or financial assistant may prefer higher recall and tolerate more flagged sentences, while a consumer search assistant may prefer higher precision to avoid alert fatigue. A deployed system should recalibrate thresholds on its own traffic and log the flagged sentence, nearest source chunk, prediction score, explanation, and reviewer correction for auditing.

V. CONCLUSION AND RECOMMENDATION

This study implemented and evaluated Evidence-Chain Reliable RAG on the complete RAGTruth data available for the experiment. The pipeline used 2,965 source records, 17,790 assistant responses, and 14,289 exact-offset hallucination spans to evaluate hallucination detection at response, sentence, and character levels. The evaluation demonstrates why multi-granularity reporting is necessary: a method can perform reasonably at response-level triage while still localizing exact unsupported spans poorly.

The results support a cautious interpretation. RandomForest achieved the best predictive scores in this experiment, including case-level F1 of 0.626 and sentence-level F1 of 0.321. Proposed-ECS-Span achieved competitive scores and added deterministic provenance explanations, but it should not be described as an overall predictive winner. Its strongest contribution is interpretability: each flag is connected to a nearest evidence chunk and a support-gap reason.

For researchers, the recommendation is to report response-level, sentence-level, and character-level metrics whenever studying RAG hallucination detection. For system builders, the recommendation is to deploy a two-stage reliability interface: first, score each response sentence for evidence support; second, show the nearest evidence and provenance explanation for every flagged sentence. This workflow helps reviewers inspect the evidence chain instead of relying on a single opaque confidence number. The most important next improvement is precise token-level span localization so that supported and unsupported content within the same sentence can be separated.

REFERENCES

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 632-642). Association for Computational Linguistics.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1870-1879). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186). Association for Computational Linguistics.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated evaluation of retrieval augmented generation. arXiv. https://arxiv.org/abs/2309.15217

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023). RARR: Researching and revising what language models say, using language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning (pp. 3929-3938). PMLR.

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In International Conference on Learning Representations.

Honovich, O., Choshen, L., Aharoni, R., Neeman, E., Szpektor, I., & Abend, O. (2022). TRUE: Re-evaluating factual consistency evaluation. In Proceedings of NAACL-HLT 2022 (pp. 3905-3920). Association for Computational Linguistics.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of EMNLP 2020 (pp. 6769-6781). Association for Computational Linguistics.

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In Proceedings of SIGIR 2020 (pp. 39-48). Association for Computing Machinery.

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In Proceedings of EMNLP 2020 (pp. 9332-9346). Association for Computational Linguistics.

Laban, P., Kryscinski, W., Agarwal, D., Fabbri, A. R., Xiong, C., Joty, S., & Wu, C.-S. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics, 10, 163-177.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv. https://arxiv.org/abs/1907.11692

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of EMNLP 2023. Association for Computational Linguistics.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906-1919). Association for Computational Linguistics.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of EMNLP 2023. Association for Computational Linguistics.

Pagnoni, A., Balachandran, V., & Tsvetkov, Y. (2021). Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Proceedings of NAACL-HLT 2021 (pp. 4812-4829). Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP-IJCNLP 2019 (pp. 3982-3992). Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proceedings of KDD 2016 (pp. 1135-1144). Association for Computing Machinery.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In Proceedings of NAACL-HLT 2018 (pp. 809-819). Association for Computational Linguistics.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of NAACL-HLT 2018 (pp. 1112-1122). Association for Computational Linguistics.

Wu, Y., Zhu, J., Xu, S., Shum, K., Niu, C., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.