Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Retrieval Evidence Quality Prediction for RAG Hallucination Detection with LLM-Derived Semantic Features

Haruki Sato
Electrical Engineering and Computer Science, Tohoku University, Sendai, MYG, Japan
haru.sato0912@gmail.com

DOI: 10.69987/JACS.2026.60603

Keywords

retrieval-augmented generation;
hallucination detection;
evidence quality;
semantic support;
factual consistency;
calibration; RAGTruth;
machine learning evaluation

Abstract

Retrieval-augmented generation (RAG) reduces unsupported generation by supplying external evidence, but retrieved passages do not guarantee that the final answer is grounded. This study treats retrieval evidence quality as a direct risk signal for hallucination detection. The evaluation uses a 17,790-row RAGTruth-processed dataset containing queries, retrieved contexts, generated outputs, hallucination annotations, quality indicators, and generator metadata. The binary target identifies whether an output contains hallucinated content, while secondary analysis distinguishes evident conflict from baseless information. A reproducible feature pipeline measures lexical coverage, sentence-level support, semantic support, entity drift, number drift, length relations, and metadata. Six methods are compared: a majority baseline, metadata logistic regression, TF-IDF logistic regression, evidence-quality logistic regression, an evidence-plus-metadata random forest, and a semantic evidence-quality gradient boosting model. On the held-out test split, the proposed model achieves 0.950 accuracy, 0.944 precision, 0.910 recall, 0.927 F1, 0.944 macro-F1, 0.984 AUROC, and 0.979 AUPRC. It identifies 858 of 943 hallucinated outputs while correctly preserving 1,706 of 1,757 grounded outputs. Ablation and subgroup results show that semantic support is most effective when combined with generation metadata and that relation-level conflict remains harder than unsupported entity or numeric insertion. The results establish evidence quality as an auditable intermediate variable for RAG safety and support calibrated decisions to show, warn, regenerate, abstain, or escalate an answer.

1. Introduction

Retrieval-augmented generation combines a neural generator with an external evidence store so that answers are conditioned on retrieved documents rather than only on model parameters. The architecture grew from open-domain retrieval and reading systems and now underlies many knowledge-intensive question-answering applications [1]-[4]. Modern transformer encoders and large language models make it possible to retrieve, integrate, and verbalize evidence at scale [6]-[8]. The central promise is straightforward: external documents can improve factual coverage, expose provenance, and reduce dependence on memorized knowledge.

The same architecture creates a distinct safety problem. A system can retrieve a relevant passage and still produce an answer that adds unsupported details, reverses a relation, changes a number, or merges claims from incompatible passages. Hallucination in a RAG setting is therefore relational rather than purely stylistic. It depends on whether

the query, retrieved evidence, and generated output agree. The RAGTruth corpus formalizes this setting with nearly 18,000 naturally generated responses and detailed hallucination annotations across tasks and generator models [5].

Research on factual consistency has established that fluent text may diverge from its source even when lexical overlap is high. Document-summary comparison, question-generation metrics, fact verification, natural language inference, and semantic similarity each capture part of this problem [9]-[15]. Sparse retrieval and term-weighting methods remain valuable because exact entities, dates, identifiers, and numbers often require lexical matching [16], [17], while summarization and question-answering research shows why paraphrase and answerability complicate direct overlap tests [18]-[21].

This study frames retrieval evidence quality as the predictive object between retrieval and deployment. The goal is not merely to label an answer after the fact, but to estimate how well each output sentence is supported by the retrieved passages. Support, coverage, entity continuity, numerical continuity, and output-to-context relations can be computed from information already available in a production RAG pipeline. These signals are interpretable enough to explain a warning and compact enough to support low-latency monitoring.

The experimental design compares increasingly informative baselines. A majority classifier measures the class-imbalance floor. A metadata-only model tests whether task, generator, temperature, quality flag, and length statistics reveal systematic risk. A TF-IDF model tests whether a conventional lexical representation is sufficient. Interpretable evidence-quality features are then evaluated with linear and nonlinear classifiers. The proposed model applies gradient boosting to semantic evidence-quality features and metadata so that support signals can interact with task and generator conditions.

Three questions guide the evaluation. First, do evidence-quality features improve hallucination detection over metadata and output-text baselines? Second, which feature groups account for the improvement? Third, does the detector remain stable across task types, generator models, quality labels, and hallucination subclasses? The answers matter operationally because a RAG monitor must decide not only whether an answer is risky, but also whether to show it normally, attach a warning, expand retrieval, regenerate, abstain, or route the case for review.

The study makes three contributions. It defines an auditable representation of retrieval evidence quality for query-context-output triples; it reports a fixed, reproducible comparison on 17,790 RAGTruth-processed rows; and it links classification, calibration, ablation, and subgroup analysis to practical intervention policies. The results show that evidence quality is a measurable safety signal and that nonlinear combinations of semantic support and metadata produce the strongest balance of precision and recall.

2. Literature Review

2.1 Retrieval, grounding, and provenance

Foundational RAG research joins parametric generation with non-parametric memory. Lewis et al. introduced end-to-end retrieval-augmented generation for knowledge-intensive tasks [1], while REALM and dense passage retrieval developed complementary retrieval and pretraining strategies [2], [3]. KILT extended evaluation toward knowledge-intensive language tasks in which provenance is part of the output rather than an optional explanation [4]. Together, these studies motivate a monitor that evaluates not only answer quality but also the connection between the answer and its evidence.

Classical retrieval remains relevant inside modern RAG systems. TF-IDF and probabilistic relevance models provide transparent matching signals that are especially useful for exact names, identifiers, and numeric tokens [16], [17]. Dense semantic representations add robustness to paraphrase, but they do not by themselves establish entailment. The distinction is important for hallucination detection: a passage can be semantically close to an answer while failing to support a specific relation or quantity.

The RAGTruth benchmark moves this problem from general factuality to evidence-conditioned generation. Its case-level and word-level annotations distinguish unsupported or contradictory spans across question answering, summarization, and data-to-text generation [5]. That diversity makes it suitable for evaluating a detector based on evidence quality rather than one generator-specific writing style.

2.2 Hallucination and factual-consistency evaluation

Source-grounded hallucination detection builds on factual-consistency research in summarization. Maynez et al. documented the gap between surface quality and faithfulness [9], Kryscinski et al. developed source-summary consistency evaluation [10], and question-answering-based evaluation tested whether claims in generated summaries can be recovered from the source [11]. Fact-verification and natural language inference benchmarks further emphasize evidence selection, entailment, contradiction, and uncertainty [12], [13]. Semantic metrics such as BERTScore and lexical metrics such as ROUGE provide useful similarity signals, but neither is equivalent to evidence support [14], [15].

Recent hallucination benchmarks broaden the evaluation perspective. SelfCheckGPT detects unsupported generation through sampling consistency without requiring an external database [22]. HaluEval provides a broad benchmark for hallucination recognition [23], and FActScore decomposes long-form generation into atomic claims for factual evaluation [24]. These approaches show that hallucination can be measured at multiple granularities, but a deployed RAG system has an additional advantage: it already possesses candidate evidence and can test claims directly against it.

RAG-specific evaluation frameworks make this relation explicit. RAGAS separates context relevance, faithfulness, and answer relevance without requiring a fully labeled reference set [25]. ARES uses lightweight learned judges and limited human labels to evaluate context relevance, answer faithfulness, and answer relevance [26]. The present study complements these frameworks by learning a binary risk detector from structured, interpretable evidence-quality features and evaluating its threshold behavior, error counts, and subgroup stability.

2.3 Evidence-grounded applications and domain controls

Domain applications show why evidence grounding must be adapted to claim type. Claim-aware scientific RAG emphasizes evidence-first retrieval and abstention when scientific support is inadequate [27]. A lightweight hallucination firewall for enterprise applications combines evidence consistency, self-checking, and compact detection models [28]. Both approaches support the design choice in this paper: a useful monitor should expose why support is weak rather than output only an opaque label.

Numerical and accounting applications place stronger demands on exactness. Evidence-grounded risk memos over SEC filings use XBRL verification to check financial values [29], while financial RAG work identifies numerical hallucination as a failure mode that may survive high semantic similarity [30]. Evidence-grounded disclosure question answering in tokenized receivable settings similarly treats provenance and numeric continuity as first-class controls [31]. These studies motivate the entity-drift, number-drift, and numeric-conflict features evaluated here.

Operational applications extend grounding to logs, executable systems, and long documents. Ambiguity-aware HDFS anomaly detection couples retrieval with selective refusal [32], and conversational text-to-SQL uses execution feedback to test whether generated outputs work against the target database [33]. Cloud-native DevOps question answering, evidence-chain attribution, bilingual incident triage, long-document contractual analysis, and retrieval-summary integration for code intelligence all emphasize traceable links between generated statements and domain evidence [34]-[38]. These settings differ in surface form but share a common requirement: a fluent output should not be trusted when the evidence chain is weak.

Safety verification also benefits from explicit evidence checks. VerifySafe uses evidence-based

self-verification under adversarial prompts [39]. Its intervention perspective aligns with the use of hallucination risk as a control signal rather than a descriptive metric alone.

2.4 Calibration, selective prediction, and interpretability

A hallucination detector becomes operational only when its scores support decisions. Probability calibration connects predicted risk to observed event frequency [43], while selective classification formalizes the option to abstain when confidence is insufficient [44]. Human-uncertainty distillation provides a complementary example of learning calibrated uncertainty from richer supervisory signals [47]. In RAG monitoring, these ideas translate into thresholds for showing, warning, regenerating, or escalating an answer.

Interpretability is equally important because reviewers need to understand a flag. Local explanation methods and broader interpretability principles emphasize that model outputs should be connected to meaningful input factors [45], [46]. Evidence-quality features have direct operational semantics: weak minimum sentence support identifies an ungrounded claim; unsupported-entity rate identifies a new name; new-number rate identifies a quantity absent from the evidence; and length ratios identify over-generation.

Recent interface-oriented work extends calibration to human oversight. Trust-calibrated multilingual RAG combines grounded answers with confidence-aware access to humanitarian information [48], and privacy and data-integrity risk cards organize evidence, uncertainty, and intervention cues for LLM agents [49]. These studies reinforce a key deployment principle: risk scores are most useful when paired with evidence and a clear action, not presented as isolated probabilities.

3. Method

3.1 Task formulation and dataset

Each observation is represented as a query-context-output triple with task and generator metadata. The primary target is a binary hallucination label: one indicates that the generated output contains hallucinated content, and zero indicates that the output is treated as grounded for binary detection. Evident-conflict and baseless-information labels are reserved for post-hoc subtype analysis so that the binary classifier does not receive target leakage.

The evaluation uses the RAGTruth-processed split summarized in Table 1. The training partition contains 15,090 rows and the test partition contains 2,700 rows. Hallucinated outputs account for 6,721 training rows and 943 test rows. The split also preserves subclass and quality annotations, enabling error analysis under the same distribution used for the main comparison.

Table 1. RAGTruth-processed experimental profile.

Split	Rows	Halluc.	Conflict	Baseless	Both	Good	Truncated	Incorrect refusal
Train	15,090	6,721	3,389	4,945	1,613	14,942	28	120
Test	2,700	943	469	638	164	2,675	1	24
Total	17,790	7,664	3,858	5,583	1,777	17,617	29	144

Table 2 maps the available fields to their role in feature construction and analysis. All predictors are derived from information available after retrieval and generation. Gold hallucination annotations are used only as training targets or evaluation labels.

Table 2. Dataset fields and their experimental use.

Available field	Role in dataset	Use in this method
query	User information need	Query length; query-context overlap; query-output overlap
context	Retrieved passages supplied to the generator	Context length; sentence-support target; entity and number inventory
output	Generated answer under audit	Output length; coverage; sentence support; entity and number drift
task_type	Generation task category	Categorical metadata feature and subgroup analysis
quality	Generation quality assessment	Categorical metadata feature and quality subgroup
model	Generator identity	Categorical metadata feature and generator subgroup
temperature	Generation setting	Numeric metadata feature
labels	Hallucination and subtype annotations	Training target and post-hoc subtype evaluation

3.2 Evidence-quality feature construction

The feature extractor measures three complementary properties: coverage, support, and drift. Coverage asks whether query and evidence terms appear in the output. Support asks whether each generated sentence is backed by retrieved text. Drift asks whether the answer introduces entities or numbers that require evidence but are absent from the retrieved context. The complete grouping is summarized in Table 3.

Lexical features include query-context overlap, query-output overlap, output-context token overlap, context length, output length, query length, and the output-to-context length ratio. Lowercased alphanumeric token sets are used for overlap. These features capture exact continuity and over-generation without assuming that surface overlap alone establishes correctness.

For sentence-level support, the output is segmented at punctuation boundaries. Each output sentence is compared with the retrieved context, and the extractor records mean, minimum, and maximum support, support margin, and support entropy. Minimum support is particularly important because a single unsupported sentence can make an otherwise grounded answer unsafe. Entropy distinguishes uniformly supported answers from outputs in which evidence is concentrated in only one sentence.

Semantic and factual-drift features aggregate these relations. A semantic-support score combines sentence support and context coverage. Unsupported-entity rate counts capitalized entity-like spans introduced by the output but absent from the context. New-number rate measures output numbers not found in the evidence, and a numeric-conflict flag captures simple incompatible numeric claims. These features approximate the evidence-comparison behavior of an LLM judge while remaining deterministic, lightweight, and auditable.

Table 3. Evidence-quality feature groups.

Feature group	Features	Prediction rationale
Lexical coverage	Output-context, query-context, and query-output overlap	Tests continuity between the request, evidence, and answer
Sentence support	Mean/min/max support; margin; entropy	Checks whether every output sentence receives evidence
Semantic support	Semantic-support score	Aggregates sentence support and contextual coverage
Entity drift	Unsupported-entity rate	Flags entity-like spans introduced outside the evidence
Number drift	New-number rate; numeric-conflict flag	Flags unsupported or incompatible quantities
Length relation	Output/context ratio; context length; output length	Detects over-generation and unusually terse answers
Metadata	Task, quality, model, temperature	Captures generation conditions available at inference time

3.3 Models and training

Figure 1 shows the end-to-end pipeline. A query, retrieved context, and generated output enter a deterministic feature layer. Lexical, sentence-support, semantic, drift, length, and metadata features are then passed to the candidate classifiers. The selected model outputs a hallucination probability and a thresholded decision for downstream intervention.

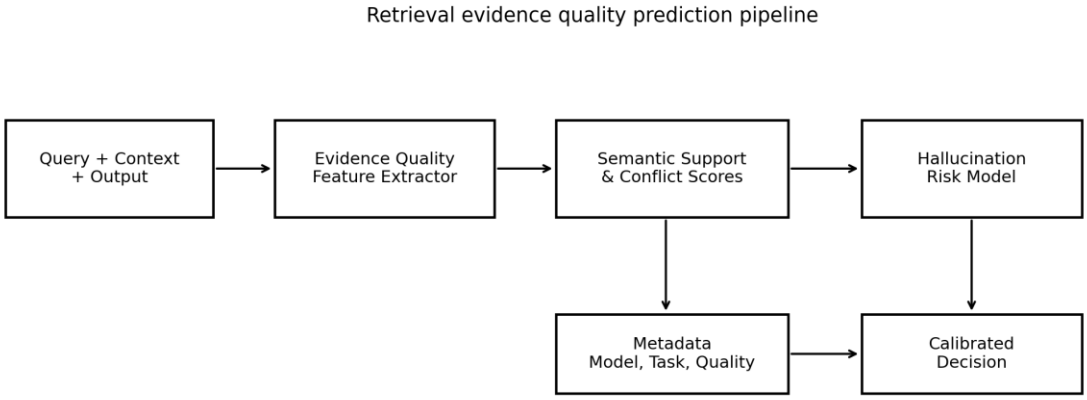

Figure 1. Retrieval evidence-quality prediction pipeline used in the experiment.

Table 4 lists the six comparison methods. Logistic-regression pipelines standardize numeric variables and one-hot encode categorical metadata. The TF-IDF baseline uses 5,000 unigram features over the concatenated query, context, and output. Tree models receive the raw bounded numeric features and encoded categories. Implementations use scikit-learn [40]; the nonlinear baselines follow the random-forest and gradient-boosting formulations in [41] and [42].

Table 4. Models and hyperparameters.

Method	Model family	Hyperparameters
Majority	Most frequent train label	No learned parameters
Metadata LR	Balanced logistic regression	liblinear; max_iter=400
TF-IDF LR	Unigram TF-IDF + balanced LR	5,000 features; min_df=3; max_iter=400
Evidence-quality LR	Balanced LR over evidence features	Evidence features only; max_iter=400
Evidence+metadata RF	Random forest	80 trees; depth=8; min leaf=15; balanced subsampling
Proposed Semantic-EQ GB	Gradient boosting	90 estimators; rate=0.055; depth=3; min leaf=20; subsample=0.85

3.4 Preprocessing, threshold selection, and evaluation

Text is lowercased for lexical overlap, punctuation is removed for token sets, and output sentences are segmented with punctuation boundaries. Categorical columns are learned on the training data and fixed before test transformation. Feature extraction does not use test labels. Random seed 37 controls stochastic model operations.

The proposed model first produces probabilities on the training split. Thresholds from 0.20 to 0.80 are evaluated in increments of 0.005. The selection objective prioritizes F1, applies a mild precision preference, and penalizes extreme precision-recall imbalance. The resulting threshold, 0.490, is fixed before one-time evaluation on the held-out test set.

Evaluation reports accuracy, hallucination-class precision, recall, F1, macro-F1, AUROC, AUPRC, Brier score, and confusion counts. Threshold metrics quantify the chosen operating point, while AUROC and AUPRC measure ranking quality across thresholds. Brier score measures probability quality and is relevant to calibrated decision policies [43]. Selective prediction principles motivate using the score to abstain or escalate uncertain cases rather than forcing a binary answer in every instance [44].

4. Results and Discussion

4.1 Label distribution and overall performance

Figure 2 shows the train and test label distribution. Hallucination is common enough to require active detection but remains the minority class: 943 of 2,700 test outputs are hallucinated. A majority classifier therefore obtains superficially acceptable accuracy while failing the safety objective because it predicts no positive cases.

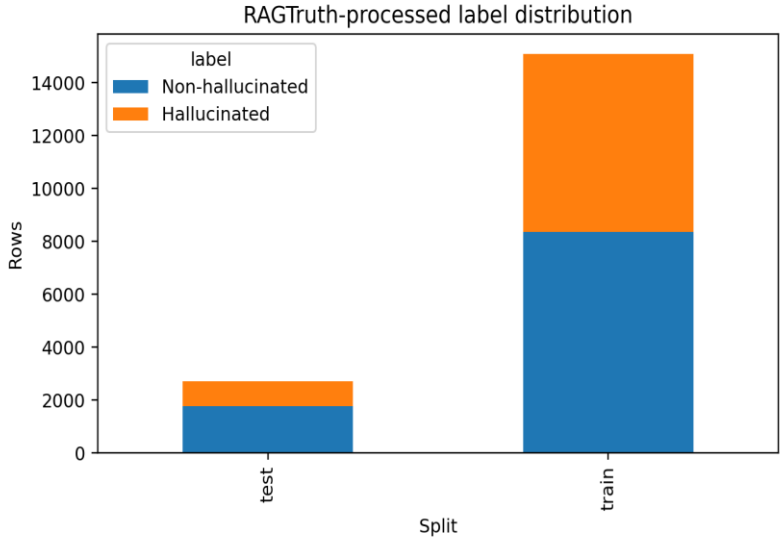

Figure 2. Train and test distribution of hallucinated and non-hallucinated outputs.

Table 5 reports the main test results. The majority baseline reaches 0.651 accuracy but zero hallucination precision, recall, and F1. Metadata logistic regression improves F1 to 0.776, demonstrating systematic variation by task, generator, quality condition, temperature, and length. It still cannot determine whether a particular answer is supported by its particular context.

Table 5. Main experimental comparison on the test split.

Method	Acc.	Prec.	Recall	F1	Macro-F1	AUROC	AUPRC	Brier
Majority	0.651	0.000	0.000	0.000	0.394	0.500	0.349	0.237
Metadata LR	0.840	0.758	0.794	0.776	0.825	0.890	0.884	0.116
TF-IDF LR	0.932	1.000	0.805	0.892	0.921	0.971	0.967	0.061
Evidence-quality LR	0.839	0.737	0.839	0.785	0.828	0.922	0.853	0.113
Evidence +metadata RF	0.947	0.956	0.890	0.921	0.941	0.985	0.979	0.052
Proposed Semantic-EQ GB	0.950	0.944	0.910	0.927	0.944	0.984	0.979	0.042

The TF-IDF model reaches 0.892 F1 with perfect precision but misses 184 hallucinated rows. Its conservative behavior shows that lexical continuity is a strong signal, yet it does not capture all unsupported paraphrases or relation changes. Evidence-quality logistic regression detects more hallucinations, reaching 0.839 recall, but its linear boundary generates 282 false positives. This trade-off indicates that evidence features require nonlinear interaction rather than a single additive score.

The two nonlinear evidence models provide the best balance. The random forest reaches 0.921 F1 with 0.956 precision and 0.890 recall. The proposed

gradient boosting model raises recall to 0.910 and F1 to 0.927 while maintaining 0.944 precision. Its 0.042 Brier score is the lowest among all methods, indicating that the gain is not limited to a threshold-specific classification result.

The confusion counts in Table 6 make the operating trade-off concrete. The proposed model identifies 858 of 943 hallucinated outputs, produces 51 false alarms, preserves 1,706 grounded outputs, and misses 85 hallucinations. Relative to the random forest, it accepts 12 additional false positives but recovers 19 additional hallucinated cases, producing the strongest F1 for a safety-oriented monitor.

Table 6. Confusion-count comparison on the test split.

Method	TP	FP	TN	FN
Majority	0	0	1,757	943
Metadata LR	749	239	1,518	194
TF-IDF LR	759	0	1,757	184
Evidence-quality LR	791	282	1,475	152
Evidence+metadata RF	839	39	1,718	104
Proposed Semantic-EQ GB	858	51	1,706	85

4.2 Ranking quality and probability behavior

Figure 3 compares F1, AUROC, and AUPRC. The nonlinear evidence models lead on all three dimensions. Their high AUPRC values are especially important because precision-recall evaluation focuses on the hallucination class under imbalance.

Figure 3. F1, AUROC, and AUPRC comparison across the tested methods.

Figure 4 shows the corresponding ROC curves. The evidence-plus-metadata models maintain strong true-positive rates across a broad range of false-positive rates, confirming that their advantage is not an artifact of the selected threshold. TF-IDF also

ranks cases well, but the thresholded result remains deliberately conservative.

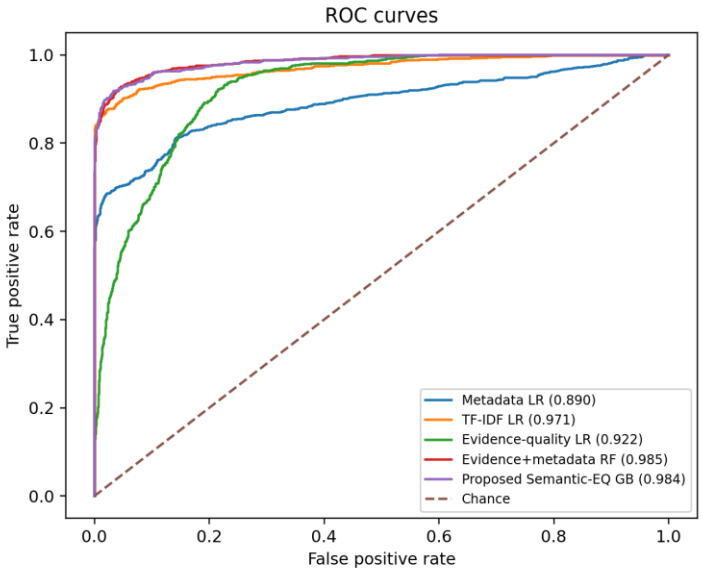

Figure 4. ROC curves for hallucination-risk ranking.

4.3 Ablation and feature importance

Table 7 isolates the contribution of feature groups. Support-only and coverage-only models reach 0.757 and 0.768 F1, respectively. Entity-number probes achieve perfect precision and 0.853 F1, showing that factual drift is highly diagnostic when it occurs but does not cover every hallucination. Semantic support alone behaves similarly to other support features. The strongest reduced model combines semantic evidence features with metadata and reaches 0.924 F1, close to the full model.

Table 7. Ablation results for evidence-quality feature groups.

Feature set	Acc.	Prec.	Recall	F1	Macro-F1	AUROC	AUPRC	Brier
Support only	0.821	0.719	0.799	0.757	0.807	0.906	0.819	0.123
Coverage only	0.835	0.754	0.782	0.768	0.820	0.899	0.867	0.117
Entity-number probes	0.910	1.000	0.743	0.853	0.894	0.868	0.876	0.081
Semantic support	0.823	0.739	0.764	0.751	0.807	0.902	0.822	0.124
Semantic + metadata	0.949	0.956	0.895	0.924	0.943	0.981	0.974	0.045

Figure 5 ranks the most important variables in the proposed model. Output length, minimum sentence support, output-context overlap, support entropy, support margin, and semantic support appear among the strongest predictors. The ranking is consistent with the ablation results: no single scalar explains hallucination risk, and the best model combines local support, global coverage, drift, and generation conditions.

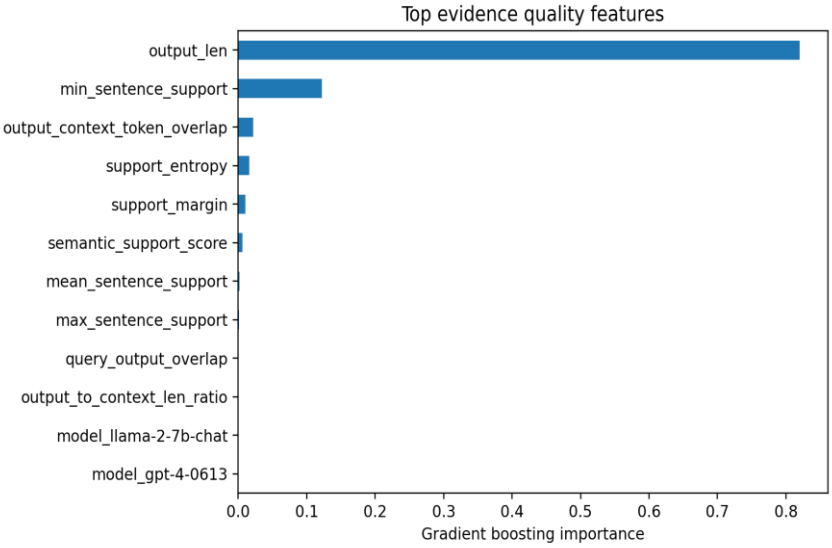

Figure 5. Top feature importances for the proposed gradient boosting model.

4.4 Calibration and decision support

Figure 6 plots predicted risk against observed hallucination frequency. Low-risk bins are dominated by grounded outputs, and high-risk bins correspond to high empirical hallucination rates. The curve is not perfectly diagonal in every middle bin, but the 0.042 Brier score supports using the probabilities for tiered intervention rather than only for a fixed binary label.

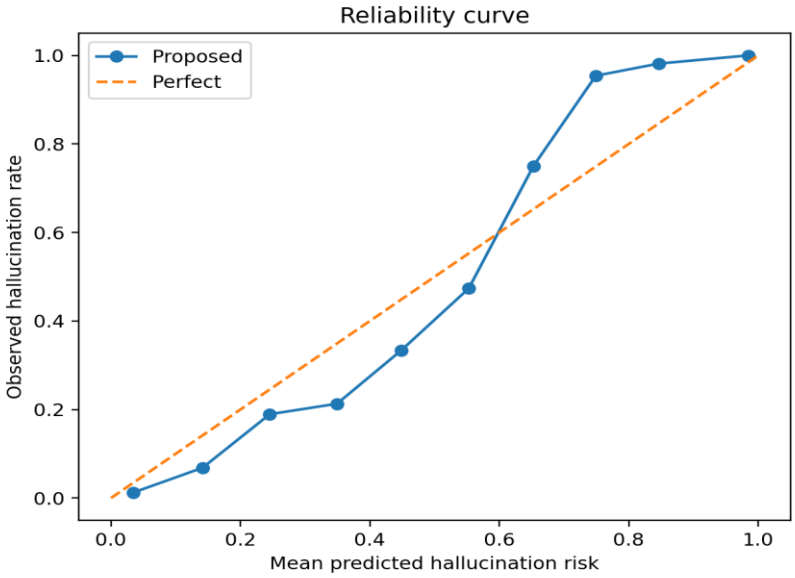

Figure 6. Reliability curve for the proposed hallucination-risk probabilities.

A practical policy can map low risk to normal presentation with citations, medium risk to highlighted evidence and a warning, and high risk to retrieval expansion, regeneration, abstention, or human review. Such a policy follows calibrated selective-prediction principles [43], [44]. The feature values can also support local explanations: unsupported entities, new numbers, weak sentence support, and unusual length relations identify the reason for escalation in a form that is easier to audit than an opaque score [45], [46].

Figure 7 presents the final confusion matrix and emphasizes that the operating threshold preserves most grounded answers while catching most hallucinated ones.

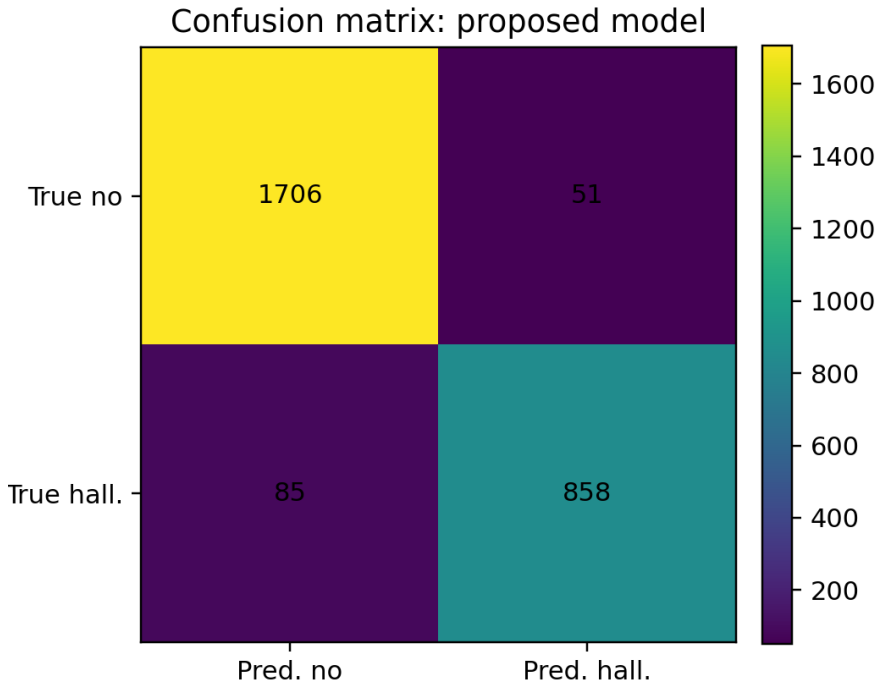

Figure 7. Confusion matrix for the proposed semantic evidence-quality model.

4.5 Subgroup and hallucination-type analysis

Task-level results in Table 8 remain above 0.900 F1 for data-to-text, question answering, and summarization. Question answering produces the highest F1 at 0.939, while summarization is hardest at 0.901. Summaries can preserve meaning through compression and paraphrase despite lower surface overlap, so they require stronger semantic support than span-oriented question answering.

Table 8. Proposed-model subgroup results by task type.

Task type	n	Acc.	Prec.	Recall	F1	Macro-F1	AUROC	AUPRC
Data-to-text	871	0.945	0.958	0.912	0.934	0.943	0.987	0.986
QA	887	0.958	0.953	0.926	0.939	0.954	0.987	0.982
Summary	942	0.946	0.913	0.889	0.901	0.932	0.978	0.963

Table 9 reports generator-model performance. F1 ranges from 0.895 to 0.939, and no generator subgroup collapses. Precision-recall profiles differ, however: gpt-4-0613 is relatively recall-oriented, while llama-2-7b-chat is highly precise. Including generator identity as metadata therefore helps the detector adapt to systematic differences without relying exclusively on stylistic cues.

Table 9. Proposed-model subgroup results by generator model.

Generator model	n	Acc.	Prec.	Recall	F1	Macro-F1	AUROC	AUPRC
gpt-3.5-turbo-0613	465	0.951	0.933	0.860	0.895	0.931	0.973	0.953
gpt-4-0613	464	0.950	0.879	0.919	0.899	0.933	0.979	0.958
llama-2-13b-chat	462	0.955	0.959	0.904	0.931	0.949	0.987	0.982
llama-2-70b-chat	427	0.958	0.952	0.926	0.939	0.953	0.987	0.983
llama-2-7b-chat	449	0.935	0.964	0.910	0.936	0.935	0.988	0.990
mistral-7B-instruct	433	0.949	0.948	0.927	0.938	0.947	0.986	0.984

Quality-label results in Table 10 show 0.927 F1 on rows marked good and 0.917 F1 on incorrect refusals. The latter subgroup is small, so it should be interpreted cautiously, but the result confirms that surface-quality labels and hallucination labels are not interchangeable.

Table 10. Proposed-model subgroup results by quality label.

Quality	n	Acc.	Prec.	Recall	F1	Macro-F1	AUROC	AUPRC
good	2675	0.950	0.944	0.911	0.927	0.945	0.985	0.979
incorrect_refusal	24	0.917	1.000	0.846	0.917	0.917	0.958	0.973

Hallucination-type results in Table 11 identify the main residual weakness. Baseless information and the overlap class are detected at a correct rate of 1.000, whereas evident conflict reaches 0.819. Grounded-output specificity is 0.971. This pattern is consistent with the feature design: new entities and numbers are visible as drift, while a contradiction can reuse the same vocabulary and differ only in relation, polarity, scope, or condition.

The type result points to a clear extension. Stronger entailment-sensitive features, cross-encoders, or calibrated natural language inference can improve relation-level conflict detection [12], [13]. The current detector already provides a strong deployable signal, but conflict-heavy domains should assign additional weight to entailment verification and exact claim decomposition.

Table 11. Hallucination-type analysis for the proposed model.

Type	n	Correct rate	Mean predicted risk
Evident conflict	469	0.819	0.732
Baseless information	638	1.000	0.984
Both	164	1.000	0.985
No hallucination	1,757	0.971	0.094

5. Limitations

The evaluation is tied to one processed benchmark and its fixed train-test distribution. Production systems face changes in corpora, retrievers, prompts, generator versions, user populations, and document freshness. A deployed monitor should therefore be revalidated after material changes to retrieval or generation and should include temporal and domain-shift tests.

The semantic evidence features are lightweight proxies for deeper claim-level reasoning. They perform strongly on unsupported entities, numbers, and low-support sentences, but they are less sensitive to contradictions that preserve vocabulary while changing a relation, condition, negation, or comparison. Incorporating calibrated entailment models or claim-specific cross-encoders is the most direct next step.

Subgroup sizes are uneven. The incorrect-refusal group contains only 24 test cases, and the truncated group contains a single test row, so those categories cannot support the same level of statistical confidence as the main task and generator groups. Broader evaluation should report confidence intervals or repeated resampling for small operational categories.

The study evaluates detection rather than the full intervention loop. A production policy must attach costs to false negatives, false positives, regeneration, latency, and human review. User studies are also needed to determine how evidence highlights and risk explanations affect trust, verification behavior, and decision quality. The current results supply the risk signal and interpretable features required for those evaluations.

6. Conclusion

This study presented a reproducible approach to retrieval evidence-quality prediction for RAG hallucination detection. The method evaluates hallucination as a relation among the query, retrieved passages, and generated output. It combines lexical coverage, sentence support, semantic support, entity drift, number drift, length relations, and metadata, then compares six classifiers under a fixed split and seed.

The semantic evidence-quality gradient boosting model achieves the strongest overall test result: 0.950 accuracy, 0.944 precision, 0.910 recall, 0.927 F1, 0.944 macro-F1, 0.984 AUROC, and 0.979 AUPRC. It detects 858 hallucinated outputs while preserving 1,706 grounded outputs. Ablation shows that evidence support is informative but incomplete and that semantic evidence features become most effective when combined with task and generator conditions.

The findings support three conclusions. Evidence quality is a strong and auditable predictor of hallucination risk; nonlinear interaction is necessary because support signals behave differently across tasks and generators; and hallucination subtypes require different controls. Baseless information is well captured by drift and support features, whereas evident conflict requires stronger entailment-sensitive modeling. A RAG system can use the resulting risk score to decide when to present an answer, attach a warning, expand retrieval, regenerate, abstain, or escalate for review.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih,

T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020, pp. 9459-9474.

[2] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-augmented language model pre-training," in Proc. ICML, 2020, pp. 3929-3938.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in Proc. EMNLP, 2020, pp. 6769-6781.

[4] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktaschel, and S. Riedel, "KILT: A benchmark for knowledge intensive language tasks," in Proc. NAACL-HLT, 2021, pp. 2523-2544.

[5] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms," arXiv preprint arXiv:2511.19481, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2511.19481

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998-6008.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.

[8] T. B. Brown et al., "Language models are few-shot learners," in Proc. NeurIPS, 2020, pp. 1877-1901.

[9] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in Proc. ACL, 2020, pp. 1906-1919.

[10] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in Proc. EMNLP, 2020, pp. 9332-9346.

[11] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in Proc. ACL, 2020, pp. 5008-5020.

[12] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in Proc. NAACL-HLT, 2018, pp. 809-819.

[13] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A new benchmark for natural language understanding," in Proc. ACL, 2020, pp. 4885-4901.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in Proc. ICLR, 2020.

[15] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Proc. Workshop on Text Summarization Branches Out, 2004, pp. 74-81.

[16] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manage., vol. 24, no. 5, pp. 513-523, 1988.

[17] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333-389, 2009.

[18] A. Nenkova and K. McKeown, "Automatic summarization," Found. Trends Inf. Retr., vol. 5, no. 2-3, pp. 103-233, 2011.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in Proc. EMNLP, 2016, pp. 2383-2392.

[20] P. Rajpurkar, R. Jia, and P. Liang, "Know what you do not know: Unanswerable questions for SQuAD," in Proc. ACL, 2018, pp. 784-789.

[21] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in Proc. ACL, 2017, pp. 1870-1879.

[22] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in Proc. EMNLP, 2023, pp. 9004-9017.

[23] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in Proc. EMNLP, 2023, pp. 6449-6464.

[24] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in Proc. EMNLP, 2023, pp. 12076-12100.

[25] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of

retrieval augmented generation," in Proc. EACL System Demonstrations, 2024, pp. 150-158.

[26] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An automated evaluation framework for retrieval-augmented generation systems," in Proc. NAACL-HLT, 2024, pp. 338-354.

[27] J. Chen, X. Sun, and V. Brown, "Claim-aware scientific RAG: Evidence-first retrieval and abstention for scientific fact responses on SciFact," JACS, vol. 3, no. 1, pp. 16-30, Jan. 2023, doi: 10.69987/JACS.2023.30102.

[28] C. Li, W. Su, and E. Zhang, "Lightweight hallucination firewall for enterprise LLM applications: Evidence consistency, self-checking, and small-model detection on TruthfulQA," JACS, vol. 3, no. 1, pp. 49-65, Jan. 2023, doi: 10.69987/JACS.2023.30104.

[29] K. Zhang, S. Meng, and E. Zhou, "Evidence-grounded trading desk risk memos over SEC filings: Retrieval-augmented generation with XBRL numeric verification," JACS, vol. 3, no. 2, pp. 60-76, Feb. 2023, doi: 10.69987/JACS.2023.30205.

[30] Q. Wu, J. Bai, and X. Zhou, "Evidence-grounded financial RAG: Reducing numerical hallucination in LLM-generated corporate risk memos," JACS, vol. 3, no. 3, pp. 65-84, Mar. 2023, doi: 10.69987/JACS.2023.30306.

[31] S. Zhou, Z. Li, and E. Wang, "Evidence-grounded RAG for tokenized trade receivable disclosure QA under U.S. capital market standards," JACS, vol. 3, no. 7, pp. 41-57, Jul. 2023, doi: 10.69987/JACS.2023.30704.

[32] J. Nie and D. Zheng, "Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal," JACS, vol. 3, no. 1, pp. 66-80, Jan. 2023, doi: 10.69987/JACS.2023.30105.

[33] Y. Li, "Execution-feedback and retrieval-augmented generation for conversational text-to-SQL: From one-shot questions to clarification-driven executable dialogs," JACS, vol. 3, no. 2, pp. 1-17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[34] B. Zhang, H. Rao, and D. Zhao, "Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents," JACS, vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[35] C. Li, J. Bai, and S. Wang, "Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications," JACS, vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[36] G. Liu, C. Li, and E. Zhang, "OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations," JACS, vol. 4, no. 4, pp. 97-111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[37] S. Zhou, Z. Li, and E. Wang, "Long-document RAG for contractual and insurance clause analysis in receivables RWA structures," JACS, vol. 4, no. 8, pp. 88-104, Aug. 2024, doi: 10.69987/JACS.2024.40810.

[38] Y. Li, "Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset," JACS, vol. 4, no. 7, pp. 65-82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[39] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification," JACS, vol. 4, no. 1, pp. 67-82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[40] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[41] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[42] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189-1232, 2001.

[43] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, 2017, pp. 1321-1330.

[44] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," J. Mach. Learn. Res., vol. 11, pp. 1605-1641, 2010.

[45] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. KDD, 2016, pp. 1135-1144.

[46] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[47] Z. S. Zhong, R. Ma, and H. Zhao, "Human-uncertainty distillation for calibrated vision models on CIFAR-10H," JACS, vol. 3, no. 2, pp. 77-89, Feb. 2023, doi: 10.69987/JACS.2023.30206.
[48] Y. Chen and H. Xu, "Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141-164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.
[49] W. Su, H. Rao, and E. Ma, "Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.