

Evaluating Evidence Sensitivity and Robustness of Hallucination Detectors in Retrieval- Augmented Generation

Ramana Kulanthaivelu

kulanthaiveluramana@gmail.com

Research Article

Keywords:

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11400108/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Retrieval-Augmented Generation (RAG) improves the accessibility of external knowledge for large language models, but generated responses can remain unsupported, contradictory, or otherwise hallucinatory. Conventional hallucination-detector evaluation largely emphasizes aggregate benchmark performance, which may not reveal whether detector decisions remain reliable when the evidence distribution changes. This study presents a controlled evaluation of hallucination detectors from an evidence-sensitivity and robustness perspective. Three detector configurations - TF-IDF with Logistic Regression, TF-IDF with Linear Support Vector Machine, and DeBERTa-v3-base were evaluated using RAGTruth, HaluBench, and TRIVIA+. In-domain evaluation on RAGTruth showed that DeBERTa achieved an AUROC of 0.8483 and Macro-F1 of 0.7616. However, performance declined substantially under distribution shift, with AUROC decreasing to 0.5348 on HaluBench and 0.5910 on TRIVIA+. Controlled evidence interventions showed substantial detector responsiveness, with mean Evidence Sensitivity Index values of 0.2641 for evidence removal and 0.3268 for cross-example evidence substitution. A context-complexity stress test further showed progressive degradation as irrelevant evidence increased from 0% to 100% within a fixed 512-token observation window, with AUROC decreasing from 0.8306 to 0.7325 and Expected Calibration Error increasing from 0.1375 to 0.4072. Input ablation demonstrated that both generated answers and retrieved evidence contribute to detection performance. These findings show that benchmark accuracy alone is insufficient to characterize hallucination-detector reliability and motivate evaluation frameworks that jointly assess generalization, evidence sensitivity, contextual robustness, and calibration.

1. INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language generation, enabling applications in question answering, summarization, information retrieval, education, scientific assistance, and decision-support systems. However, they can also produce responses that are fluent and coherent while containing factually incorrect, unsupported, or contextually inconsistent information. Such failures, commonly referred to as hallucinations, remain a major challenge for the reliable deployment of generative artificial intelligence. Existing hallucination-detection research has explored a range of signals, including internal model representations, uncertainty estimation, response consistency, reasoning behaviour, and semantic characteristics by Lokesh et al., (2026); Mukkawar et al., (2026); Zhang et al., (2026); Liu et al., (2024); Joo et al., (2025); Kang et al., (2026).

Retrieval-Augmented Generation (RAG) reduces reliance on static parametric knowledge by providing retrieved external documents or passages to the generative model as contextual evidence. Ideally, generated responses should therefore be grounded in the retrieved information. However, retrieval does not eliminate hallucination: a model may ignore relevant evidence, misinterpret it, combine supported and unsupported information, or generate claims that are not supported by the retrieved context. The RAGTruth corpus provides substantial empirical evidence of this problem and has become an important benchmark for studying hallucination detection in RAG systems Saha et al., (2026); Niu et al., (2024).

This has shifted research attention from improving retrieval and generation alone toward determining whether generated content is actually supported by retrieved evidence. Existing approaches include sentence- and token-level verification, uncertainty estimation, self consistency analysis, semantic analysis, and evidence-aware detection Mukkawar et al., (2026); Saha et al., (2026); Venkatachalapathy, (2026); R et al., (2026); Zheng et al., (2025). However, strong performance on conventional classification metrics does not necessarily establish that a detector is responding appropriately to the relationship between a generated answer and its supporting evidence. Accuracy, precision, recall, F1-score, and AUROC measure predictive performance, but they provide limited information about whether a detector's decision remains responsive when the evidence underlying that decision changes Lokesh et al., (2026); Mukkawar et al., (2026); Zhang et al., (2026); Liu et al., (2024); Kou et al., (2024); Gu et al., (2024).

This concern is reinforced by recent evidence on generalisation. Dubanowska et al., (2025) showed that representation-based hallucination detectors can exploit spurious correlations and experience substantial degradation under out-of-distribution evaluation. Consequently, strong performance on a familiar benchmark may partly reflect properties of the training distribution, generation process, or dataset construction rather than robust detection of hallucination itself. Recent evidence-grounded approaches further emphasise the importance of the detector–evidence relationship. RLSeek by Huang et al., (2026) demonstrated the value of explicitly grounding verification in retrieved evidence, while Evidence-Aligned Entity Verification by Jia et al., (2026) investigated evidence alignment and stability under evidence perturbation. These developments motivate evaluation procedures that examine how detector predictions respond when the evidence associated with a generated answer is systematically modified.

The evaluation setting itself can also influence detector reliability. TRIVIA + by Chen et al., (2026) highlights limitations in existing hallucination benchmarks through long-context RAG examples and controlled label-noise conditions, demonstrating that conventional held-out evaluation may not fully characterize detector behaviour. In this study, contextual complexity is examined more specifically by progressively increasing the proportion of irrelevant evidence within a fixed 512-token detector observation window. This design isolates the effect of evidence dilution from changes in the maximum sequence length and provides a controlled means of assessing prediction stability and calibration.

Motivated by these limitations, this study evaluates hallucination detectors from an evidence sensitivity and robustness perspective. It combines conventional in-domain evaluation with cross-dataset distribution-shift testing and controlled evidence interventions involving evidence removal, cross-example evidence substitution, and the addition of irrelevant distractor evidence. Detector responsiveness to these interventions is quantified using an Evidence Sensitivity Index and prediction-flip rate, while calibration and contextual robustness are evaluated alongside conventional classification metrics. The study compares classical text based detectors with a transformer-based detector and examines whether strong benchmark performance is accompanied by consistent evidence-dependent behaviour. Through this unified evaluation, the study aims to provide a more diagnostic assessment of hallucination detector reliability beyond aggregate benchmark accuracy.

2. RELATED WORK

2.1. LLM Hallucinations

Hallucination in LLMs is generally characterized by generated content that is factually incorrect, unsupported, inconsistent with available evidence, or otherwise unreliable despite being linguistically plausible. The phenomenon is not necessarily uniform and may arise from different failure mechanisms. The broader taxonomy reviewed by Liu et al., (2024) distinguishes factuality-based, faithfulness-based, logical, and hybrid forms of hallucination and emphasizes that hallucination can emerge through interactions between data, training, model behaviour, and inference processes. Liu et al., (2024) also argues that conventional evaluation can provide insufficient insight into the underlying mechanisms of hallucination.

Recent work has further demonstrated that hallucination may occur at different levels of reasoning. MathHalu by Zhang et al., (2026), for example, separates hallucinations in mathematical reasoning into **cognitive hallucinations**, associated with failures in understanding or interpreting the problem, and **logical hallucinations**, associated with invalid reasoning or inference. Experimental results reported by Zhang et al., (2026) indicate that current reasoning-capable models may still confuse these categories and experience substantial difficulties in self-detection. These findings reinforce the idea that hallucination cannot necessarily be represented adequately by a single undifferentiated binary label.

Detection research has therefore explored a wide range of observable signals. Attention-based approaches investigate internal structural properties of model attention in Lokesh et al., (2026), while uncertainty-based research examines token entropy, semantic uncertainty, and prediction variability in Mukkawar et al., (2026); Joo et al., (2025); Kang et al., (2026). Such approaches provide useful signals but have different assumptions regarding model accessibility, computational cost, interpretability, and transferability.

2.2. RAG Systems and Hallucinations

Retrieval-Augmented Generation was introduced to reduce the limitations associated with relying exclusively on parametric model knowledge by incorporating external information into the generation process. However, retrieval does not guarantee that the generated response is faithful to the retrieved information. RAGTruth by Niu et al., (2024) was developed specifically to address this problem and contains nearly 18,000 naturally generated RAG responses with manual annotations at both the response and word levels. The corpus demonstrates that unsupported and contradictory claims may persist even when relevant evidence is supplied to the model.

The importance of evidence grounding has motivated systems that integrate verification directly into RAG pipelines. HALT-RAG (Saha et al., 2026), for example, combines sentence level attribution, verification, and token-level hallucination detection and uses evidence support to determine whether generated responses should be accepted, modified, re-retrieved, or refused. The approach illustrates a

shift from conventional generation toward explicit verification of generated content. Other application-specific work has similarly incorporated retrieval into hallucination detection. Reference from Zheng et al., (2025) integrates RAG with model fine-tuning for factual hallucination detection in an educational setting, while Zhong et al., (2026) investigates self consistency-based detection in cybersecurity and highlights the need for domain-specific hallucination detection in high-risk environments.

These studies establish the importance of external evidence but also expose an important question for detector evaluation: a system can have access to evidence without necessarily using that evidence correctly.

2.3. Hallucination Detection Methods

Hallucination-detection approaches can broadly be divided into several categories.

2.3.1. Internal model signals

Reference from Lokesh et al., (2026) extracts structural and statistical features from attention maps and combines spectral characteristics with measures related to uncertainty and variability. The approach demonstrates that internal representations can provide signals that are not directly available from surface-level text classification.

2.3.2 Uncertainty and consistency signals

Reference from Mukkawar et al., (2026) combines semantic uncertainty, self-consistency, and claim decomposition and reports that these signals provide complementary information. In that study, self-consistency produced the strongest individual signal, while combining multiple indicators improved overall ranking performance. Reference from Joo et al., (2025) similarly explores sentence-level entropy and shows that sentence-level uncertainty can differentiate some hallucinated and non-hallucinated outputs more effectively than token-level entropy, while also documenting examples that remain difficult to detect.

Reference from Kang et al., (2026) provides a broader examination of uncertainty quantification for hallucination detection and notes important limitations, including the difficulty of interpreting uncertainty scores, the computational cost of model ensembles, temporal changes in factual knowledge, and the challenge of applying uncertainty estimates consistently to long-form generation.

2.3.3 Reasoning-oriented approaches

MathHalu by Zhang et al., (2026) demonstrates that hallucination detection becomes more complicated when hallucination occurs within a reasoning process rather than only in the final answer. The reported confusion between cognitive and logical hallucinations also indicates that a detector may correctly identify an unreliable output while still failing to identify the underlying failure type. These approaches collectively demonstrate that hallucination detection is not limited to simple text classification. However,

they also motivate evaluation methods that can determine whether detected signals correspond to the actual evidence supporting a generated claim.

2.4. Transformer-Based Detection

Transformer-based models provide a natural foundation for supervised hallucination detection because they can jointly encode questions, retrieved contexts, and generated responses. The literature supplied in this study includes transformer-based detection using models such as GPT-Neo and TinyLlama with downstream classifiers by Lokesh et al., (2026), as well as larger families of standard and reasoning-oriented LLMs evaluated directly on hallucination detection tasks by Zhang et al., (2026). Other work has adopted pre-trained transformer architectures for hybrid detection frameworks and evidence verification. However, increasing model capacity does not automatically resolve the reliability problem. A detector may learn correlations associated with particular datasets or generation procedures instead of learning general properties of hallucination. This concern becomes particularly important under distribution shift, where performance on a familiar benchmark may not reflect behaviour in a new environment.

2.5. Evidence-Grounded Detection

Evidence-grounded detection represents an important development beyond purely output based hallucination detection. Rather than asking only whether an answer appears uncertain or inconsistent, evidence-grounded methods explicitly examine whether generated content can be supported by retrieved material. HALT-RAG by Saha et al., (2026) integrates evidence-aware generation with sentence-level verification and token-level hallucination detection. More recent research strengthens this direction. RLSeek by Huang et al., (2026) reports that reasoning based hallucination detectors can produce incorrect predictions when verification chains lack explicit grounding in source evidence and therefore introduces explicit evidence quotation during verification. EAEV by Jia et al., (2026) aligns generated entities with retrieved evidence and introduces counterfactual stability under evidence perturbation as part of its verification framework.

These developments are particularly relevant to the present work. They indicate that **evidence is not merely contextual information but an object whose relationship with the generated claim can itself be evaluated**. However, existing studies primarily use evidence alignment as a component of their detection or verification framework. The present study instead places **detector responsiveness to controlled evidence changes** at the centre of the evaluation methodology.

2.6. Robustness and Domain Shift

A major limitation of conventional hallucination evaluation is the assumption that benchmark performance transfers to unseen data. Reference from Dubanowska et al., (2025) provides strong evidence against this assumption. Dubanowska et al. (2025) found that representation-based hallucination detectors evaluated on RAGTruth can exploit spurious correlations and that analysed methods performed close to random under out-of-distribution evaluation. This finding is especially important because a detector intended for real-world RAG deployment is likely to encounter changes in,

- document sources
- domains
- generating models
- question styles
- context structure
- retrieval distributions.

Reference from R et al., (2026) similarly highlights the importance of calibration and cross-lingual robustness when evaluating hybrid hallucination detectors, while Gu et al., (2024) argues that existing hallucination evaluation remains fragmented across horizontal comparisons between models and domains and vertical analysis of individual reasoning processes.

Consequently, out-of-distribution testing should be viewed not simply as an additional benchmark score but as a means of testing whether a detector has learned generalizable hallucination characteristics.

2.7. Context Complexity and Irrelevant-Evidence Robustness

Long-context generation creates an additional challenge because the presence of more evidence does not necessarily imply better evidence use. As contextual length increases, relevant information can be surrounded by increasing amounts of irrelevant or weakly related content, potentially affecting both generation and hallucination detection.

Reference from Kang et al., (2026) identifies long-form generation as a limitation for uncertainty based detection because a single uncertainty score may not adequately represent the correctness of individual claims within a multi-claim response. It therefore motivates finer-grained evaluation.

More recently, TRIVIA + by Chen et al., (2026) was specifically introduced to address the lack of long-context RAG hallucination benchmarks. The benchmark contains long-context examples and controlled noisy-label variants, and its evaluation shows that current detectors remain below a potential performance ceiling while label noise can degrade detection performance.

TRIVIA+ therefore provides an important environment for **stress testing** rather than a gap that remains unexplored. In the present study, this issue is investigated through a controlled context complexity stress test in which the proportion of irrelevant evidence is increased within a fixed 512-token observation window

2.8. Research Gap and Positioning of the Present Study

The reviewed literature shows that hallucination detection has progressed from simple output level classification toward uncertainty-based, internal-state, reasoning-based, and evidence grounded approaches. Existing work has also demonstrated significant challenges involving distribution shift, long-context evaluation, label noise, benchmark artifacts, and evidence grounding Liu et al., (2024); Gu et al.,

(2024); Kang et al., (2026); Dubanowska et al., (2025); Chen et al., (2026). Recent evidence-grounded methods such as RLSeek and EAEV further demonstrate that explicit interaction with retrieved evidence can improve hallucination detection and robustness

However, these strands of research leave an important evaluation issue that motivates this study. **A detector's classification performance does not, by itself, establish that the detector is responding to evidence validity.** A model could produce a correct prediction for the original example while relying on linguistic patterns, response style, model-specific characteristics, or benchmark-specific correlations. Conversely, a detector that genuinely performs evidence verification should exhibit predictable changes in its confidence or classification when evidence is removed, replaced with evidence from another example, or supplemented with irrelevant distractor content.

Accordingly, the central focus of this study is not simply to develop another hallucination classifier, but to **evaluate evidence dependence as a measurable property of hallucination detectors.** The proposed framework combines conventional in-domain evaluation with out-of-distribution testing, controlled evidence interventions, evidence-dilution stress testing, and calibration analysis. This design is intended to determine not only whether a detector can classify hallucinations, but also **whether it behaves consistently with the evidence-grounded nature of the task it is intended to solve.**

3. METHODOLOGY

3.1. Research Questions

This study investigates the reliability of hallucination detectors in retrieval-augmented generation (RAG) under in-domain evaluation, distribution shift, controlled evidence modification, and increasing contextual complexity. The evaluation is guided by four research questions:

RQ1

How effectively do conventional machine-learning and transformer-based models detect hallucinations in retrieval-augmented generation under in-domain evaluation?

RQ2

How does hallucination-detection performance change when models are evaluated under distribution shifts that differ from the training benchmark?

RQ3

To what extent are hallucination detectors sensitive to controlled changes in retrieved evidence, including evidence removal, cross-example evidence substitution, and the addition of irrelevant distractor evidence?

RQ4

How does increasing the proportion of irrelevant evidence within a fixed detector observation window affect hallucination-detection performance, evidence sensitivity, prediction stability, and calibration?

These questions deliberately separate conventional classification performance from evidence-dependent reliability. RQ1 establishes baseline detection capability, RQ2 evaluates transfer beyond the training distribution, RQ3 examines detector responsiveness to explicit evidence interventions, and RQ4 measures robustness to increasing irrelevant contextual information.

3.2. Experimental Design

The study adopts a controlled multi-stage evaluation design rather than proposing a new hallucination-generation or retrieval system. The central objective is to evaluate whether a trained hallucination detector remains reliable when the information available to it changes.

RAGTruth is used as the primary development benchmark. A context-grouped split is first constructed from the RAGTruth training data, with groups defined by retrieved context, to prevent the same context from appearing in both model-development subsets. Two conventional text-based classifiers, TF-IDF with Logistic Regression and TF-IDF with Linear Support Vector Machine (SVM), are trained as classical baselines. A pretrained DeBERTa-v3-base transformer is fine-tuned as the transformer-based detector.

The resulting detectors are evaluated on the held-out RAGTruth test set for RQ1. For RQ2, the trained detectors are evaluated without retraining or threshold adjustment on HaluBench and TRIVIA+ test data. HaluBench is additionally evaluated after excluding examples whose source_ds is RAGTruth, providing a more conservative external evaluation.

RQ3 uses the trained DeBERTa detector as a fixed model and applies controlled evidence interventions to otherwise unchanged RAGTruth test examples. RQ4 further evaluates DeBERTa under controlled evidence-composition changes in which the proportion of irrelevant evidence increases while the detector's total observation window remains fixed. Finally, an input-ablation analysis compares Answer-only, Question+Answer, Evidence+Answer, Question+Evidence, and Question+Evidence+Answer configurations to identify the relative contribution of the available information sources.

No external test set is used for model training, threshold optimization, or hyperparameter selection.

3.3. Datasets

3.3.1. RAGTruth

RAGTruth is the primary dataset for detector development and in-domain evaluation. The processed version (Weights and Biases, n.d.) used in this study was obtained from the Weights & Biases Hugging Face repository, wandb/RAGTruth-processed. The repository provides the RAGTruth records with fields

including the query, retrieved context, generated output, task type, model metadata, and processed hallucination labels.

The downloaded data contain 15,090 training examples and 2,700 test examples. Each example contains a question/query, retrieved context, generated answer/output, and hallucination annotations. For this study, the response is converted to a binary hallucination label: an example is assigned label 1 when either the processed `evident_conflict` or `baseless_info` indicator is equal to 1, and label 0 otherwise.

The RAGTruth training set contains 8,369 non-hallucinated and 6,721 hallucinated examples, corresponding to a hallucination proportion of approximately 44.5%. The held-out test set contains 1,757 non-hallucinated and 943 hallucinated examples, corresponding to approximately 34.9%.

3.32. HaluBench

HaluBench is used as an external distribution-shift benchmark. The dataset was introduced by (Ravi et al., 2024). The published dataset contains 14,900 test examples with passage, question, answer, binary label, and source-dataset metadata. The complete downloaded HaluBench set contains 7,730 non-hallucinated and 7,170 hallucinated examples. Its `source_ds` field identifies the underlying source benchmark. Because HaluBench includes 900 examples identified as originating from RAGTruth, two evaluations are reported: the complete 14,900 example HaluBench set and a conservative 14,000-example subset obtained by excluding all examples with `source_ds=RAGTruth`.

3.3.3. TRIVIA+

TRIVIA+ is used as a second external distribution-shift benchmark. It was introduced by (Chen, Padmanabhan, Giyahchi, et al., 2026) in *Rethinking Evaluation for LLM Hallucination Detection: A Desiderata, A New RAG-based Benchmark, New Insights*, published in the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics in 2026. The benchmark was specifically designed to address deficiencies in existing hallucination-detection benchmarks, including the scarcity of RAG-based evaluation with long contexts and the lack of realistic label-noise stress testing.

The processed TRIVIA+ data used in this study contain 3,223 usable examples after excluding one record with a missing answer. The official split structure contains training, validation, and test partitions. Only the test partition, consisting of 645 examples, is used for external evaluation in this study. The TRIVIA+ test set contains 421 non-hallucinated and 224 hallucinated examples.

3.4. Dataset summary

Table 1
Dataset Summary

Dataset	Role	Examples used	Hallucination rate
RAGTruth train	Model development	15,090	44.5%
RAGTruth test	In-domain evaluation	2,700	34.9%
HaluBench	External evaluation	14,900	48.1%
HaluBench excluding RAGTruth-source	Conservative external evaluation	14,000	50.1%
TRIVIA+ test	External evaluation	645	34.7%

3.5. Preprocessing

3.5.1. Label harmonization

Because the three datasets use different annotation schemas, their labels are mapped to a common binary representation:

$$y = \begin{cases} 1, & \text{hallucination} \\ 0, & \text{non hallucination} \end{cases}$$

For RAGTruth, label 1 is assigned when either `evident_conflict` or `baseless_info` is present in the processed response-level annotation. For HaluBench, the published PASS/FAIL label is mapped to the binary task label. TRIVIA+ response-level labels are mapped to the same representation.

3.5.2. Duplicate and overlap checks

Potential data leakage was assessed before model training. Exact normalized-context overlap between the RAGTruth training and test partitions was zero. A context-grouped development split was also constructed so that no retrieved context occurred simultaneously in its training and validation subsets.

Cross-dataset checks found no exact context overlap between RAGTruth training data and HaluBench. Exact Question+Context+Answer overlap was also absent. In addition, the dedicated 14,000-example HaluBench analysis excludes the 900 records whose source metadata explicitly identifies RAGTruth as their origin.

3.5.3 Transformer input construction

The DeBERTa detector receives the question, evidence, and answer as a single packed sequence. A maximum observation window of 512 tokens is used. To ensure that neither the question nor the generated answer is completely removed through truncation, separate budgets of 128 tokens are allocated to the question and answer. The remaining token capacity is allocated to retrieved evidence,

yielding an effective evidence budget of up to 252 tokens after accounting for the beginning-of-sequence and end-of-sequence tokens.

The packed structure is therefore conceptually represented as

[BOS] + Q + E + A + [EOS]

where **Q**, **A**, and **E** denote the question, answer, and evidence components, respectively.

This 512-token representation is treated as the detector's fixed observation window. Consequently, the context-complexity experiments investigate changes in the **composition** of the observation window rather than increasing the maximum sequence length.

3.5.4. Classical model representation

For the conventional baselines, question, evidence, and answer are concatenated into a textual representation:

X=[Q; A; E]

A TF-IDF representation using unigram and bigram features is then constructed from the RAGTruth development data. The vectorizer is fitted only on the model-development data and subsequently applied unchanged to all test and external datasets.

3.6. Data Splitting

The 15,090 RAGTruth training examples are partitioned into model-development training and validation sets using retrieved-context grouping. This produces 13,578 training examples and 1,512 validation examples, with zero context overlap between the two subsets.

The RAGTruth test set contains 2,700 examples and is not used during model fitting, hyperparameter selection, or threshold selection. The test set is also context-disjoint from the RAGTruth training data.

The development validation set is used during DeBERTa training for epoch-level model selection using Macro-F1. The final trained detector is subsequently evaluated on the untouched RAGTruth test set and the external benchmarks.

For cross-dataset evaluation, HaluBench and TRIVIA + are used strictly as external evaluation sets. No external test labels are used to modify the trained models or their operating threshold.

3.7. Models

3.7.1 TF-IDF + Logistic Regression

TF-IDF features are combined with a binary Logistic Regression classifier. The vectorizer uses unigram and bigram features, sublinear term frequency scaling, minimum document frequency filtering, and a maximum feature count of 200,000. Logistic Regression is trained using **C = 1.0** and a maximum of 2,000 iterations.

3.7.2. TF-IDF + Linear SVM

A linear Support Vector Machine is trained using the same TF-IDF representation as the Logistic Regression baseline. The classifier uses **C = 1.0** and a maximum of 5,000 iterations.

The SVM's decision function is used directly for classification and ranking. A sigmoid transformation is applied only when probability-like values are required for certain numerical diagnostics. These transformed values are **not treated as calibrated probabilities**.

3.7.3. DeBERTa-v3-base

The transformer detector uses microsoft/deberta-v3-base with a newly initialized binary sequence-classification head. The model is fine-tuned on the context-grouped RAGTruth training partition for three epochs using a maximum sequence length of 512 tokens. The learning rate is 2×10^{-5} , weight decay is 0.01, and gradient norm clipping is applied with a maximum norm of 1.0. Gradient accumulation is used during GPU training.

The model is selected using validation Macro-F1, with the best validation checkpoint retained for subsequent experiments. Mixed-precision training is used during GPU execution while the model parameters remain in FP32 before automatic mixed-precision processing.

Importantly, the same trained DeBERTa model is retained as a fixed detector throughout the evidence-intervention, context-complexity, and input-ablation experiments; these experiments do not involve retraining.

3.8. Metrics

Performance is evaluated using complementary threshold-dependent, threshold-independent, and probabilistic reliability measures.

- **Accuracy** measures the proportion of correctly classified examples.
- **Precision, recall, and F1-score** quantify performance for the hallucination class, while **Macro-F1** gives equal weight to the hallucination and non-hallucination classes.
- **AUROC** measures the ability of the detector score to rank hallucinated examples above non-hallucinated examples independently of a particular classification threshold.
- **AUPRC** measures ranking performance in precision-recall space and is particularly informative when the positive-class prevalence changes between datasets.
- **Brier score** measures the squared difference between predicted hallucination probability and the binary outcome:

$$Brier = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2$$

,

- For DeBERTa, calibration is additionally evaluated using **Expected Calibration Error (ECE)** with ten probability bins. Lower ECE indicates closer agreement between predicted probabilities and empirical outcome frequencies. Because the Linear SVM does not generate calibrated probabilities, its sigmoid-transformed decision scores are not interpreted as calibrated posterior probabilities.

3.9. Evidence Interventions

RQ3 evaluates whether detector predictions change when the retrieved evidence associated with an otherwise unchanged question-answer pair is modified.

Four evidence conditions are evaluated using the fixed DeBERTa detector.

3.9.1 Original evidence

The original RAGTruth question, retrieved evidence, and generated answer are provided:

(Q, E, A) ,

This produces the baseline hallucination probability $(P(H|E))$.

3.9.2. Evidence removal

The retrieved evidence is removed while retaining the original question and answer:

$(Q, \emptyset, A)$,

The resulting prediction is compared with the original prediction.

3.9.3. Cross-example evidence substitution

The original evidence is replaced by a retrieved context taken from a different RAGTruth test example while the question and answer remain unchanged:

$(Q, E_j, A), j \neq i$,

This condition is described as **cross-example evidence substitution** rather than guaranteed semantic irrelevance because the substituted context is not manually validated for semantic unrelatedness in every individual case.

3.9.4. Distractor addition

A second context from another RAGTruth test example is appended to the original evidence: $(Q, E_i \oplus E_j, A)$

This condition introduces irrelevant or competing contextual material while retaining the original evidence.

3.9.5. Evidence Sensitivity Index

For each example, the Evidence Sensitivity Index (ESI) is defined as

$$ESI_i = |P_i(H|\bar{E}) - P_i(H|E)|$$

where E is the original evidence and $\bar{E}$ represents an intervention condition.

Higher ESI indicates a larger change in hallucination probability following the evidence intervention.

Prediction-flip rate is also calculated as the proportion of examples for which the binary prediction changes relative to the original condition.

3.10. Context Complexity

RQ4 evaluates the effect of increasing irrelevant information while keeping the detector's maximum observation window fixed at 512 tokens.

Five evidence-composition conditions are evaluated:

- 0% irrelevant evidence: 100% original evidence
- 25% irrelevant evidence: 75% original + 25% distractor
- 50% irrelevant evidence: 50% original + 50% distractor
- 75% irrelevant evidence: 25% original + 75% distractor
- 100% irrelevant evidence: distractor evidence only

For every example, the question and answer remain unchanged. The evidence region retains a fixed token budget, so the experimental factor is the **proportion of relevant versus irrelevant evidence within the observation window**, rather than the total maximum sequence length.

For each condition, the detector is evaluated using AUROC, AUPRC, Macro-F1, Brier score, ECE, mean predicted hallucination probability, ESI relative to the 0% condition, and prediction-flip rate.

This design therefore provides a controlled stress test of **context complexity and evidence dilution** without claiming to evaluate arbitrary sequence lengths beyond the detector's 512 token window.

3.11. Input Ablation

An input-ablation experiment is performed to investigate which information components contribute most strongly to detector decisions. The same RAGTruth test examples are evaluated under five configurations:

1. **A**: Answer only
2. **Q + A**: Question and answer
3. **E + A**: Evidence and answer
4. **Q + E**: Question and evidence
5. **Q + E+A**: Question, evidence, and answer

The same trained DeBERTa model is used for all conditions. No additional training is conducted for any ablation.

Performance is compared using accuracy, precision, recall, F1, Macro-F1, AUROC, AUPRC, Brier score, and ECE. In addition, the change in predicted hallucination probability and prediction-flip rate relative to the full $Q + E+A$ configuration are calculated for each ablated condition. The ablation is intended to distinguish evidence-dependent information from answer dependent and question-dependent signals and to determine whether high detection performance requires the complete input configuration.

3.12. Statistical Analysis

Statistical analysis is designed to evaluate both aggregate performance changes and paired per example prediction changes.

For evidence interventions and input ablations, paired differences are calculated for each test example because the same question-answer pair is evaluated under multiple conditions. The **Wilcoxon signed-rank test** is used to test whether the distribution of paired prediction changes differs systematically from zero. For classifier comparisons on the same RAGTruth test examples, **McNemar's exact test** is used to evaluate differences in paired binary decisions.

For the context-complexity experiment, changes across the five predefined evidence-composition conditions are evaluated descriptively using AUROC, AUPRC, Macro-F1, Brier score, ECE, mean ESI, and prediction-flip rate. Paired per-example comparisons with the 0% irrelevant-evidence condition are evaluated using the Wilcoxon signed-rank test. Nonparametric bootstrap resampling with 5,000 repetitions is used to estimate 95% confidence intervals for mean intervention effects and ESI values. Bootstrap resampling is performed at the example level. Statistical significance is assessed using a two-sided significance level of $\alpha = 0.05$. Extremely small numerical p -values are reported using inequality notation rather than as zero. Because the context-complexity experiment contains five predefined levels, trend analyses are interpreted as evidence of monotonic association across the controlled conditions rather than as a substitute for a full continuous dose-response model.

4. RESULTS

4.1. RQ1 - In-Domain Detection

RQ1 evaluates the ability of classical text-based detectors and a transformer-based detector to identify hallucinated responses under in-domain conditions. All models were evaluated on the same 2,700-example RAGTruth test set, which was not used during model training or model selection.

Table 2 summarizes the results. The TF-IDF + Logistic Regression model achieved an accuracy of 0.7707, Macro-F1 of 0.7406, AUROC of 0.7988, and AUPRC of 0.7307. The TF-IDF + Linear SVM obtained an accuracy of 0.7011, Macro-F1 of 0.6893, AUROC of 0.7975, and AUPRC of 0.7290. DeBERTa-v3-base achieved an accuracy of 0.7789, Macro-F1 of 0.7616, AUROC of 0.8483, and AUPRC of 0.7786. Its Brier score was 0.1617, compared with 0.1736 for Logistic Regression and 0.2032 for the Linear SVM.

Table 2
In-domain detection performance on the RAGTruth test set

Model	Accuracy	Precision	Recall	F1	Macro-F1	AUROC	AUPRC
TF-IDF + Logistic Regression	0.7707	0.6938	0.6151	0.6521	0.7406	0.7988	0.7307
TF-IDF + Linear SVM	0.7011	0.5553	0.7243	0.6286	0.6893	0.7975	0.7290
DeBERTa-v3-base	0.7789	0.6680	0.7296	0.6974	0.7616	0.8483	0.7786

The ROC and precision-recall curves in Fig. 2 and Fig. 3 show that DeBERTa provides stronger ranking performance across decision thresholds than the two classical baselines. The AUROC difference is particularly evident, with DeBERTa reaching 0.8483 compared with approximately 0.798 for both TF-IDF-based classifiers.

The confusion matrix for DeBERTa contains 1,415 true negatives, 342 false positives, 255 false negatives, and 688 true positives. Thus, the detector identifies a substantial proportion of both classes while maintaining a higher hallucination recall than Logistic Regression.

Paired McNemar tests further indicate statistically significant differences in the binary predictions between the models. For DeBERTa versus Logistic Regression, the exact McNemar test gives $p = 2.27 \times 10^{-18}$, while comparison of DeBERTa with Linear SVM gives $p = 4.79 \times 10^{-16}$. The difference between the two classical classifiers is also statistically significant ($p = 1.22 \times 10^{-71}$). These tests establish that the models do not merely produce different continuous scores but also make systematically different binary decisions on the same test examples.

Overall, the in-domain results demonstrate that the trained DeBERTa detector provides strong discrimination on the benchmark on which it was developed. However, these results alone do not establish whether the detector generalizes to other distributions or whether its decisions are appropriately dependent on retrieved evidence.

4.2. RQ2 - Distribution Shift

RQ2 investigates how detector performance changes when evaluated on datasets that differ from the RAGTruth training environment. No model was retrained and no external test-set labels were used to modify the classifiers.

Table 3 presents the external results. On HaluBench, DeBERTa achieves an AUROC of 0.5348 and Macro-F1 of 0.5204, compared with 0.8483 and 0.7616, respectively, on the RAGTruth test set. This corresponds to an AUROC decrease of 0.3135, or 36.96%, relative to the in-domain reference. Its AUPRC decreases from 0.7786 to 0.5091, while its Brier score increases from 0.1617 to 0.2847.

A more conservative HaluBench evaluation excluding the 900 examples identified as originating from RAGTruth produces an AUROC of 0.5032 and Macro-F1 of 0.5069. Relative to the RAGTruth reference, the AUROC decrease is 40.69%. The result indicates that the reduction in performance cannot be attributed solely to the inclusion of HaluBench examples explicitly identified as originating from RAGTruth.

On TRIVIA+, DeBERTa achieves an AUROC of 0.5910 and Macro-F1 of 0.4738. Its AUROC therefore decreases by 30.34% relative to RAGTruth. The Brier score increases to 0.3283, indicating substantially poorer probabilistic reliability under this external distribution.

Table 3
Distribution-shift performance of the evaluated detectors.

Model	Dataset	Macro-F1	AUROC	AUPRC
TF-IDF + Logistic Regression	HaluBench	0.3711	0.4725	0.4590
TF-IDF + Linear SVM	HaluBench	0.4027	0.4669	0.4605
DeBERTa-v3-base	HaluBench	0.5204	0.5348	0.5091
TF-IDF + Logistic Regression	HaluBench excl. RAGTruth	0.3596	0.5008	0.5014
TF-IDF + Linear SVM	HaluBench excl. RAGTruth	0.3920	0.4918	0.4977
DeBERTa-v3-base	HaluBench excl. RAGTruth	0.5069	0.5032	0.5084
TF-IDF + Logistic Regression	TRIVIA+	0.4268	0.4662	0.3229
TF-IDF + Linear SVM	TRIVIA+	0.4612	0.4489	0.3136
DeBERTa-v3-base	TRIVIA+	0.4738	0.5910	0.4293

The classical models also exhibit substantial distribution-shift degradation. Logistic Regression decreases from an in-domain AUROC of 0.7988 to 0.4725 on HaluBench and 0.4662 on TRIVIA+. Linear SVM decreases from 0.7975 to 0.4669 and 0.4489, respectively. The corresponding F1 reductions are particularly large for Logistic Regression, with the hallucination-class F1 falling from 0.6521 on RAGTruth to 0.0692 on HaluBench and 0.1007 on TRIVIA+.

The threshold-dependent and threshold-independent measures also show why multiple metrics are necessary. For example, DeBERTa obtains a relatively high recall of 0.7768 on TRIVIA+, but precision falls to 0.3766 and accuracy to 0.4760. Thus, its high recall does not indicate reliable overall classification under this distribution shift.

These findings demonstrate a pronounced separation between in-domain benchmark performance and external generalization. Importantly, the effect persists even when the explicitly RAGTruth-derived portion of HaluBench is excluded. Therefore, the observed degradation is consistent with substantial distribution sensitivity rather than being explained solely by the identifiable source overlap.

4.3. RQ3 - Evidence Sensitivity

RQ3 examines whether the DeBERTa detector changes its predictions when the evidence associated with an unchanged question-answer pair is manipulated.

Under the original evidence condition, DeBERTa achieves an AUROC of 0.8483 and Macro-F1 of 0.7616. Removing the evidence reduces AUROC to 0.7308 and Macro-F1 to 0.5151. Replacing the evidence with a context from another RAGTruth test example produces AUROC of 0.7332 and Macro-F1 of 0.3906. In contrast, adding a distractor context to the original evidence produces only a small performance change, with AUROC of 0.8455 and Macro-F1 of 0.7555.

Table 4
Performance under evidence interventions.

Condition	Accuracy	Macro-F1	AUROC	AUPRC	Brier
Original evidence	0.7789	0.7616	0.8483	0.7786	0.1617
Evidence removed	0.5174	0.5151	0.7308	0.6339	0.3336
Cross-example substitution	0.4219	0.3906	0.7332	0.6570	0.3932
Original + distractor	0.7719	0.7555	0.8455	0.7755	0.1664

The probability-level analysis provides a more direct measure of detector responsiveness. The mean Evidence Sensitivity Index (ESI) is 0.2641 for evidence removal, with a 95% bootstrap confidence interval of [0.2533, 0.2748]. Cross-example evidence substitution produces an even larger mean ESI of 0.3268 [0.3155, 0.3381]. In comparison, adding the distractor produces a substantially smaller mean ESI of 0.0113 [0.0101, 0.0125].

Prediction-flip rates follow the same pattern. Removing evidence changes the binary prediction for 36.37% of examples, while cross-example substitution changes 50.07% of predictions. Adding a distractor changes only 1.15% of predictions.

Table 5
Evidence Sensitivity Index and prediction changes.

Intervention	Mean ESI	95% CI	Median ESI	Prediction flip rate
Evidence removed	0.2641	[0.2533, 0.2748]	0.1522	36.37%
Cross-example substitution	0.3268	[0.3155, 0.3381]	0.2579	50.07%
Original + distractor	0.0113	[0.0101, 0.0125]	0.0000	1.15%

Paired Wilcoxon tests show that all three interventions produce statistically detectable changes in predicted hallucination probability. Evidence removal yields $p = 6.54 \times 10^{-276}$, cross-example substitution produces a value below the numerical reporting precision of the test, and distractor addition produces $p = 1.07 \times 10^{-100}$. However, statistical significance should be interpreted together with effect magnitude: the distractor condition is statistically detectable because of the large number of paired observations, but its practical effect is small, as indicated by its ESI of only 0.0113 and prediction-flip rate of 1.15%.

The results therefore indicate that the detector is responsive to major modifications of the evidence associated with an answer. At the same time, the limited response to a single appended distractor suggests that simply adding irrelevant material to otherwise intact evidence does not substantially alter the detector's decision under this intervention. This distinction motivates the more controlled evidence-composition experiment in RQ4.

4.4. RQ4 - Context Complexity

RQ4 evaluates the response of the detector as the proportion of irrelevant evidence is progressively increased within the fixed 512-token observation window. The question and answer remain unchanged, while the evidence composition changes from 0% to 100% irrelevant information.

A systematic degradation is observed across all major performance measures. At 0% irrelevant evidence, DeBERTa obtains AUROC of 0.8306, AUPRC of 0.7628, and Macro-F1 of 0.7333. At 25%, these decrease to 0.8163, 0.7459, and 0.6882, respectively. At 50% irrelevant evidence, AUROC falls to 0.7833 and Macro-F1 to 0.6064. At 75%, Macro-F1 declines further to 0.4911, and at 100% irrelevant evidence it reaches 0.4044. AUROC declines to 0.7325 at the 100% condition.

Table 6
Detector performance under increasing irrelevant evidence.

Irrelevant evidence	AUROC	AUPRC	Macro-F1	Brier	ECE	Mean $P(H)$
0%	0.8306	0.7628	0.7333	0.1824	0.1375	0.4860
25%	0.8163	0.7459	0.6882	0.2118	0.1898	0.5391
50%	0.7833	0.7070	0.6064	0.2650	0.2649	0.6142
75%	0.7542	0.6819	0.4911	0.3306	0.3489	0.6982
100%	0.7325	0.6560	0.4044	0.3835	0.4072	0.7565

The deterioration is accompanied by a progressive increase in prediction instability. Mean ESI relative to the 0% condition increases from 0.0601 at 25% irrelevant evidence to 0.1362 at 50%, 0.2175 at 75%, and 0.2757 at 100%. Prediction-flip rate similarly increases from 8.70% to 20.00%, 34.41%, and 44.07%, respectively.

Calibration also deteriorates progressively. ECE increases from 0.1375 in the original condition to 0.1898, 0.2649, 0.3489, and 0.4072 as irrelevant evidence increases. At the same time, the mean predicted hallucination probability rises from 0.4860 to 0.7565, indicating a systematic shift toward higher hallucination probabilities as relevant evidence is progressively diluted.

The paired Wilcoxon comparisons against the 0% condition show significant changes at every non-zero intervention level. For 25% irrelevant evidence, $p = 4.66 \times 10^{-115}$; for 50%, $p = 1.34 \times 10^{-205}$; and for the 75% and 100% conditions the numerical test output is below the available floating-point reporting range $p < 10^{-300}$. These tests concern paired per-example prediction changes rather than the five aggregate conditions themselves.

Overall, the context-complexity experiment shows a consistent relationship between evidence dilution and detector reliability. Increasing irrelevant information within the fixed observation window is associated with lower discrimination, worse calibration, greater probability shifts, and more frequent prediction changes.

4.5. Input Ablation

The input-ablation experiment investigates which components of the Question-Evidence-Answer representation contribute to DeBERTa's predictions. All five conditions use the same trained detector and the same 2,700 RAGTruth test examples.

The Answer-only condition achieves an AUROC of 0.7512 and Macro-F1 of 0.5408, demonstrating that substantial hallucination-related signal is present in the generated answer itself. Adding the question does not improve these measures: the Question+Answer condition produces AUROC of 0.7308 and Macro-F1 of 0.5151.

In contrast, the Evidence+Answer configuration achieves AUROC of 0.8308 and Macro-F1 of 0.7439. These values are almost identical to those obtained using the full Question+Evidence+Answer configuration, which produces AUROC of 0.8305 and Macro-F1 of 0.7339.

The Question+Evidence condition performs substantially worse, obtaining AUROC of only 0.5047 and Macro-F1 of 0.4748. This indicates that, for this detector and dataset, the generated answer provides information that is critical for effective hallucination classification.

Table 7
Input-ablation results on RAGTruth test.

Input	Accuracy	Macro-F1	AUROC	AUPRC	Brier	ECE
A	0.5411	0.5408	0.7512	0.6898	0.3109	0.3287
Q + A	0.5174	0.5151	0.7308	0.6339	0.3336	0.3483
E + A	0.7596	0.7439	0.8308	0.7633	0.1759	0.1234
Q + E	0.4819	0.4748	0.5047	0.3221	0.3026	0.2316
Q + E+A	0.7470	0.7339	0.8305	0.7628	0.1820	0.1358

The paired analysis relative to the full Q + E+A configuration shows substantially larger changes when the evidence and answer components are removed. Answer-only produces a mean absolute probability difference of 0.2205 and changes the binary prediction for 31.19% of examples. Question+Answer changes 32.81% of predictions, while Question+Evidence produces the largest mean absolute probability difference, 0.3355, and changes 48.22% of predictions.

By comparison, removing only the question while retaining evidence and answer produces a mean absolute probability difference of just 0.0355 and a prediction-flip rate of 4.22%. The paired Wilcoxon test is nevertheless statistically significant $p = 1.94 \times 10^{-48}$, reflecting the large sample size and indicating that the change is detectable even though its magnitude is comparatively small.

The corresponding results for Answer-only, Question+Answer, and Question+Evidence are also statistically significant, with $p = 2.999 \times 10^{-259}$, $p = 2.685 \times 10^{-254}$, and $p = 2.611 \times 10^{-32}$, respectively.

Taken together, the ablation results indicate that the detector does not rely exclusively on retrieved evidence. The answer itself provides a substantial predictive signal, while the addition of evidence markedly improves discrimination and calibration. Conversely, removing the answer while retaining question and evidence severely reduces detection performance. The near-equivalence of Evidence+Answer and Question+Evidence+Answer in AUROC suggests that, under this experimental setting, the question contributes relatively little additional ranking information once the evidence and generated answer are available.

5. DISCUSSION

The results provide a consistent picture of hallucination detection as a reliability problem rather than a classification problem alone. Across the experiments, the DeBERTa-v3-base detector achieved strong in-domain performance on RAGTruth, but its performance deteriorated substantially under distribution shift and under controlled changes to the evidence available to the detector. The evidence-intervention and input-ablation experiments further show that the detector responds to retrieved evidence, but that its decisions are not determined by evidence alone. These findings are discussed below in relation to the four research questions.

5.1. In-Domain Detection Does Not Guarantee Robustness

The RQ1 results show that transformer-based detection provides stronger in-domain ranking performance than the two conventional TF-IDF baselines. DeBERTa achieved an AUROC of 0.8483 and AUPRC of 0.7786 on the RAGTruth test set, compared with AUROC values of 0.7988 and 0.7975 for Logistic Regression and Linear SVM, respectively. Its Macro-F1 was also higher at 0.7616.

However, the difference between the classical and transformer-based models is considerably smaller than would be suggested by model complexity alone. The TF-IDF Logistic Regression baseline achieves an AUROC close to that of the Linear SVM, while the DeBERTa improvement is more pronounced in ranking performance than in thresholded accuracy or F1. This indicates that substantial hallucination-related information is already present in surface-level textual features.

This observation is important because high in-domain performance can otherwise create an impression of detector reliability. RQ2 demonstrates that this interpretation is incomplete. The fact that DeBERTa achieves a strong RAGTruth score does not mean that the same learned representation transfers reliably to a different benchmark distribution.

5.2. Distribution Shift Reveals Substantial Detector Instability

The external evaluations provide the clearest evidence that benchmark performance does not directly translate into generalization. DeBERTa's AUROC decreased from 0.8483 on RAGTruth to 0.5348 on HaluBench and 0.5910 on TRIVIA+. The conservative HaluBench subset excluding examples identified as originating from RAGTruth yielded an AUROC of 0.5032, which is close to chance-level ranking.

The same general pattern is observed for the conventional baselines. Logistic Regression and Linear SVM also experience large decreases in AUROC and F1 when transferred to external datasets. Therefore, the observed distribution sensitivity is not unique to the transformer architecture. Rather, it appears to be a broader characteristic of supervised hallucination detection when the detector is trained on one benchmark and applied to substantially different data-generating conditions.

Several sources of distribution difference may contribute to this behaviour, including changes in domains, question formulation, evidence structure, response characteristics, source datasets, and

generation processes. The present experiments do not isolate these factors individually, and therefore the results should not be interpreted as identifying a single causal source of degradation. Nevertheless, the consistency of the degradation across multiple detectors and external datasets indicates that aggregate in-domain performance is insufficient as a standalone measure of detector robustness.

The increase in Brier score on external datasets further indicates that distribution shift affects not only classification performance but also the reliability of the model's predicted confidence. This distinction is important for practical deployment because a detector that is uncertain when it should be confident, or confident when it is unreliable, can introduce additional risk into downstream RAG systems.

5.3. Evidence Sensitivity Is Observable but Not Equivalent to Evidence Verification

RQ3 provides more direct information about how detector decisions respond to evidence. Removing evidence produced a mean ESI of 0.2641 and changed the binary prediction for 36.37% of examples. Cross-example evidence substitution produced an even larger mean ESI of 0.3268 and a prediction-flip rate of 50.07%. In contrast, appending a distractor produced a mean ESI of only 0.0113 and a prediction-flip rate of 1.15%.

These results indicate that the detector is responsive to substantial evidence changes. In particular, removing or replacing the original evidence can materially alter the predicted hallucination probability even though the question and generated answer remain unchanged. This behaviour is consistent with the detector using information from the evidence representation rather than making decisions solely from the answer text.

At the same time, the three intervention results are not equivalent. The very small effect of the distractor condition suggests that the addition of one distractor does not necessarily disrupt the detector when the original evidence remains available. This is consistent with the possibility that the detector can tolerate limited additional context when relevant evidence is still present within its observation window.

The results therefore support a more precise interpretation of evidence grounding. The detector demonstrates **evidence responsiveness**, but these experiments do not establish that it performs explicit semantic verification of every claim against every retrieved passage. In particular, cross-example substitution was not manually validated as semantically irrelevant for every sample. ESI should therefore be interpreted as a behavioural measure of sensitivity to evidence changes, rather than a direct measure of formal evidence verification.

5.4. Context Complexity Produces Progressive Degradation

RQ4 provides stronger evidence that contextual composition matters. As the proportion of irrelevant evidence increases from 0% to 100% within the fixed 512-token observation window, AUROC decreases

from 0.8306 to 0.7325 and Macro-F1 decreases from 0.7333 to 0.4044. At the same time, ECE increases from 0.1375 to 0.4072, while mean predicted hallucination probability increases from 0.4860 to 0.7565.

The approximately monotonic nature of these changes is particularly important. The detector does not merely experience a single failure under extreme context conditions; its performance and probabilistic behaviour progressively change as relevant evidence becomes a smaller component of the observation window. Prediction-flip rates similarly increase from 8.70% at 25% irrelevant evidence to 44.07% at 100%.

This suggests that irrelevant context can affect detector behaviour even when the question and answer remain unchanged. However, the experiment should be interpreted as a **context-complexity or evidence-dilution stress test**, rather than a general long-context evaluation. The model continues to operate within a fixed 512-token window throughout the experiment. Consequently, the findings concern changes in evidence composition within the detector's available observation window rather than the effects of arbitrarily increasing sequence length.

The simultaneous increase in predicted hallucination probability is also noteworthy. As relevant evidence is diluted, the detector becomes progressively more likely to classify the unchanged answer as hallucinated. This is consistent with reduced access to useful supporting evidence, although the experiment cannot distinguish between mechanisms such as evidence dilution, token-position effects, or changes in contextual interactions.

5.5. What Information Does the Detector Actually Use?

The input-ablation experiment provides an important qualification to the evidence-sensitivity results. Answer-only input achieves an AUROC of 0.7512, demonstrating that the generated answer itself contains substantial information correlated with the hallucination label. Thus, the detector is not exclusively dependent on retrieved evidence.

Adding evidence to the answer produces a large improvement: the E + A configuration reaches an AUROC of 0.8308 compared with 0.7512 for A alone. This demonstrates that evidence provides additional useful information. The full Q + E+A configuration achieves a similar AUROC of 0.8305, while Q + E performs poorly at 0.5047. The large degradation when the answer is removed indicates that the answer is an essential component of the detector's input representation.

The relatively small difference between E + A and Q + E+A suggests that, under this particular experimental configuration, the question contributes less additional ranking information once the evidence and generated answer are present. This does not mean that the question is intrinsically unimportant in hallucination detection; rather, it indicates that its marginal contribution to this detector under the RAGTruth task formulation is limited.

Taken together, the ablation and intervention findings support a nuanced interpretation: the detector is **evidence-sensitive but not evidence-exclusive**. Its decisions incorporate information from the generated

answer and retrieved evidence, with the combination producing substantially stronger performance than answer-only input.

5.6. Implications for Hallucination-Detector Evaluation

The combined results suggest that evaluation of hallucination detectors should extend beyond a single benchmark score. A detector can obtain a relatively strong AUROC on an in-domain benchmark while showing substantial performance degradation under distribution shift, changing evidence, or increasing irrelevant context.

The results therefore motivate a multidimensional evaluation protocol in which at least four properties are considered jointly: conventional classification performance, out-of-distribution generalization, evidence responsiveness, and probabilistic reliability. This framing is consistent with the broader concern in the literature that benchmark metrics alone may conceal systematic weaknesses in hallucination detection.

The experiments also suggest that future detector development should distinguish between **classification capability** and **evidence dependence**. A detector can exploit useful signals in the answer itself and still respond appropriately to changes in supporting evidence. Consequently, strong performance should not automatically be interpreted as evidence that the detector explicitly verifies every generated claim against retrieved sources.

6. LIMITATIONS AND THREATS TO VALIDITY

6.1. Fixed Transformer Observation Window

The DeBERTa detector uses a maximum 512-token observation window. The context-complexity experiment therefore evaluates evidence composition within this window rather than unrestricted long-context processing. The results should not be generalized directly to models with substantially longer context windows or architectures specifically designed for long-context reasoning.

6.2. Controlled Evidence Interventions Are Synthetic

The evidence interventions are controlled manipulations of benchmark examples rather than naturally occurring retrieval failures. Evidence removal, cross-example substitution, and distractor insertion provide useful behavioural tests but do not reproduce every form of retrieval error encountered in deployed RAG systems.

In particular, cross-example evidence substitution is not guaranteed to be semantically irrelevant for every individual example. Similarly, the distractor condition introduces one additional context, rather than systematically modelling all possible patterns of irrelevant or conflicting retrieval.

6.3. No Contradictory or Semantically Preserving Intervention

The present study does not include controlled contradictory-evidence or semantic-preserving paraphrase interventions. Consequently, the evidence-sensitivity results should not be interpreted as covering all possible forms of evidence modification. Such interventions represent important directions for future work.

6.4. Single Transformer Architecture and Training Run

Only DeBERTa-v3-base is used as the fine-tuned transformer detector, and the main transformer experiments are based on a single training run. Other architectures, model sizes, random seeds, or fine-tuning strategies may exhibit different robustness profiles. The findings should therefore be interpreted as evidence about the evaluated detector configuration rather than all transformer-based hallucination detectors.

6.5. Benchmark and Label Dependence

All experiments depend on the quality and design of the underlying benchmark annotations. Hallucination is represented as a binary label in the present experiments, although real generated responses can contain mixtures of supported and unsupported claims. Differences in annotation procedures and label distributions between RAGTruth, HaluBench, and TRIVIA + may also contribute to observed cross-dataset differences.

6.6. Distribution Shift Is Multidimensional

The external datasets differ from RAGTruth in several potentially interacting ways, including domains, question types, evidence characteristics, generation sources, and annotation processes. The experiments demonstrate performance degradation under these combined shifts but do not isolate the individual contribution of each factor.

6.7. Threshold and Calibration Considerations

F1, accuracy, precision, and recall are evaluated using a fixed decision threshold of 0.5. Under distribution shift, changes in score distributions can make this threshold suboptimal. For this reason, AUROC and AUPRC are emphasized alongside threshold-dependent metrics. The Linear SVM's sigmoid-transformed decision scores are used only as probability-like diagnostics and are not claimed to be calibrated posterior probabilities.

6.8. Baseline Representation Differences

The classical TF-IDF models and DeBERTa do not operate under identical representational constraints. The classical models use the concatenated available text representation, whereas DeBERTa uses a 512-token observation window with explicit question and answer budgets. This difference reflects the intended model configurations but should be considered when interpreting direct numerical comparisons between model families.

6.9. Packing Difference Between Canonical and Intervention Evaluations

The canonical cross-dataset evaluation and the intervention/ablation experiments use different evaluation objectives and therefore different controlled input-packing configurations. The canonical RQ1/RQ2 evaluation follows the established training representation, whereas the intervention and ablation experiments impose fixed component budgets to ensure comparability across manipulated conditions. Consequently, absolute scores from these two experiment groups should not be interpreted as identical baseline measurements. The intervention analyses focus on **within-condition changes under a common controlled representation**, while the canonical evaluation is used for the primary benchmark results.

7. CONCLUSION

This study investigated hallucination detection in retrieval-augmented generation from the perspective of **evidence sensitivity and robustness**, rather than benchmark classification performance alone. A combination of classical text-based detectors and a DeBERTa-v3-base classifier was evaluated using RAGTruth, HaluBench, and TRIVIA+, followed by controlled evidence interventions, context-complexity stress testing, and input ablation.

The in-domain results showed that DeBERTa achieved an AUROC of 0.8483 and Macro-F1 of 0.7616 on the RAGTruth test set, outperforming the evaluated TF-IDF baselines in ranking performance. However, external evaluation demonstrated substantial distribution sensitivity. DeBERTa's AUROC decreased to 0.5348 on HaluBench, 0.5032 on the HaluBench subset excluding explicitly RAGTruth-derived examples, and 0.5910 on TRIVIA+. These results show that strong performance on a familiar benchmark does not by itself establish reliable cross-dataset generalization.

The evidence-intervention experiments further demonstrated that detector decisions are responsive to substantial changes in retrieved evidence. Evidence removal produced a mean ESI of 0.2641, while cross-example evidence substitution produced a mean ESI of 0.3268. Prediction changes were also substantial, with 36.37% of predictions changing after evidence removal and 50.07% changing after evidence substitution. By comparison, adding a single distractor produced only a small mean ESI of 0.0113 and a 1.15% prediction-flip rate.

The context-complexity experiment revealed a progressive degradation as relevant evidence was diluted by irrelevant information within the fixed 512-token observation window. AUROC decreased from 0.8306 to 0.7325, Macro-F1 from 0.7333 to 0.4044, and ECE increased from 0.1375 to 0.4072 as irrelevant evidence increased from 0% to 100%. These findings indicate that detector reliability can be affected not only by whether evidence is present, but also by how relevant evidence is distributed within the available contextual representation.

Finally, input ablation showed that hallucination-related information is present in the generated answer itself, while retrieved evidence provides substantial additional information. Evidence+Answer achieved AUROC of 0.8308, compared with 0.7512 for Answer-only input, whereas Question+Evidence without the answer produced AUROC close to chance at 0.5047. Thus, the evaluated detector is best characterized as **evidence-sensitive rather than purely evidence-dependent**: it responds to changes in retrieved evidence, but also uses substantial information contained in the generated answer.

Overall, the findings support evaluating hallucination detectors along multiple dimensions rather than relying on a single benchmark score. In-domain discrimination, distribution-shift robustness, evidence sensitivity, contextual robustness, and calibration provide complementary information about detector reliability. The proposed evaluation framework offers a structured basis for identifying detectors whose benchmark performance may not fully reflect their behaviour when evidence conditions change. Future work should extend the framework to additional transformer architectures, longer observation windows, naturally occurring retrieval failures, contradictory-evidence interventions, semantic-preserving transformations, and finer-grained claim-level evaluation.

DECLARATIONS

Competing interests

The author declares that there are no relevant financial or non-financial competing interests.

Ethics approval

Not applicable. This study uses publicly available datasets and does not involve human participants or animals.

Funding

No funding was received for conducting this study or preparing this manuscript.

Author Contribution

RK: Conceptualization, methodology, data curation, software, formal analysis, investigation, visualization, validation, writing – original draft, and writing – review and editing.

Data Availability

This study used publicly available benchmark datasets and did not generate new primary data. RAGTruth was obtained from the Weights & Biases Hugging Face repository (wandb/RAGTruth-processed),

HaluBench from the Patronus AI Hugging Face repository (PatronusAI/HaluBench), and TRIVIA+ from its publicly available benchmark release. The datasets can be accessed from their respective public repositories. The specific preprocessing, dataset filtering, and evaluation procedures used in this study are described in the manuscript.

REFERENCES

1. Chen, W., Padmanabhan, V., Giahchi, T., Wong, E., & Akoglu, L. (2026). Rethinking evaluation for LLM hallucination detection: A desiderata, a new RAG-based benchmark, new insights. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14912–14931. <https://doi.org/10.18653/v1/2026.acl-long.680>
2. Dubanowska, Z., Żelaszczyk, M., Brzozowski, M., Mandica, P., & Karpowicz, M. P. (2025). Representation-based broad hallucination detectors fail to generalize out of distribution. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 17563–17575. <https://doi.org/10.18653/v1/2025.findings-emnlp.952>
3. Gu, C., Zhang, W., Huang, Z., Kou, J., Liu, Z., Zhao, C., Liu, C., Zhang, L., Lin, W., Wang, Z., Deng, J., Xie, Y., Huang, G., Zhang, C., Lu, X., Wang, C., Zhang, Z., Yuan, H., Duan, X., & Fang, Y. (2024). *LENS: Layers of evaluation of hallucination in GenAI systems*. 1–85. <https://doi.org/10.1109/uv63228.2024.11189150>
4. Huang, Z., Wen, D., Zhu, Y., Lian, X., Liang, Y., Hao, K., Li, N., Zhang, L., Zhang, Q., Wen, J.-R., Dou, Z., & Wu, F. (2026). RLSeek: Evidence-grounded reasoning for RAG hallucination detection. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 32329–32347. <https://doi.org/10.18653/v1/2026.acl-long.1492>
5. Jia, R., Fang, Z., Wu, M., Lu, J., & Zhang, Y. (2026). Evidence-aligned entity verification for hallucination detection in retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2026*, 29538–29554. <https://doi.org/10.18653/v1/2026.findings-acl.1477>
6. Joo, E., Lee, Y.-J., & Choi, H.-J. (2025). Entropy-based sentence-level hallucination score in large language models. *2025 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 77–78. <https://doi.org/10.1109/bigcomp64353.2025.00022>
7. Kang, S., Bakman, Y. F., Yaldiz, D. N., Buyukates, B., & Avestimehr, S. (2026). Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions. *IEEE BITS the Information Theory Magazine*, 1–12. <https://doi.org/10.1109/mbits.2026.3691290>
8. Kou, J., Lu, X., Yin, A., Zhang, W., Fang, X., Li, Y., Pan, T., Zhang, L., Cheng, G., Yuan, H., Duan, X., & Fang, Y. (2024). Beyond isolated fixes: A comprehensive survey on hallucination mitigation with a three-dimensional taxonomy and integrative framework. *2024 7th International Conference on Universal Village (UV)*, 1–60. <https://doi.org/10.1109/uv63228.2024.11189154>

9. Liu, Z., Kou, J., Zhang, W., Gu, C., Fang, X., Huang, Z., Yuan, H., Li, H., Lu, X., Yin, A., Liu, C., Zhao, C., Lin, W., Zhang, L., Wang, Z., Huang, H., Duan, X., & Fang, Y. (2024). *Comprehensive evaluation of AI hallucination and novel UV-Oriented framework toward safe and trustworthy AI*. 1–136. <https://doi.org/10.1109/uv63228.2024.11189137>
10. Lokesh, R., Mokshagna, C., Dinesh, K., & Usha, P. (2026). Uncertainty-aware structural attention features for hallucination detection in LLMs. *2026 International Conference on Signal Processing and Communications (SPCOM)*, 1–5. <https://doi.org/10.1109/spcom71105.2026.11622967>
11. Mukkawar, S., Pachpute, S., Paul, O., Paul, K., & Deshmukh, P. (2026). A hybrid approach for hallucination detection in large language models using semantic uncertainty, self-consistency, and claim decomposition. *2026 International Conference on Intelligent and Sustainable Electronics & Computing Technologies (INSECT)*, 1–6. <https://doi.org/10.1109/insect68872.2026.11663745>
12. Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10862–10878. <https://doi.org/10.18653/v1/2024.acl-long.585>
13. R, H., S, G., Thangavel, S. K., Rahul, B., Mahalakshmi, J., & R, S. I. (2026). HaloGuard: A practical hybrid framework for hallucination detection in retrieval-augmented generation systems : Combining multi-judge agreement, uncertainty quantification, and lexical features without embedding dependencies. *2026 International Conference on Intelligent and Sustainable AI Systems (ICOSAAS)*, 1026–1032. <https://doi.org/10.1109/icosasas68663.2026.11648754>
14. Ravi, S. S., Mielczarek, B., Kannappan, A., Kiela, D., & Qian, R. (2024). Lynx: An open source hallucination evaluation model. *arXiv Preprint*. arXiv:2407.08488
15. Saha, P., Bhattacharjee, S., & Nath, A. R. (2026). HALT-RAG: Hallucination-tolerant retrieval-augmented QA for scientific literature. *2026 International Conference on Power, Electronics, Communications, Computing, and Intelligent Infrastructure (PECCII)*, 1–6. <https://doi.org/10.1109/peccii70991.2026.11661908>
16. Venkatachalapathy, R. (2026). Designing reliable LLM systems for AI-Assisted engineering education. *2026 IEEE International Conference on Innovation, Ethics & Emerging Tech in Engineering and Computing Education (IE2C)*, 1–6. <https://doi.org/10.1109/ie2c69620.2026.11663036>
17. Weights and Biases. (n.d.). *RAGTruth-processed*. Huggingface. Retrieved October 4, 2026, from <https://huggingface.co/datasets/wandb/RAGTruth-processed>
18. Zhang, B., Gao, C., Han, B., Hu, M., Luo, Z., Zhang, J., Yao, W., Bai, X., Geng, G., & Wang, Z. (2026). MathHalu: A benchmark for mathematical reasoning process hallucination detection in large reasoning models. *ICASSP 2026–2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1426–1430. <https://doi.org/10.1109/icassp55912.2026.11460764>
19. Zheng, L., Ye, J., Zhao, G., Luo, S., Nan, M., & Xie, Y. (2025). Incorporating RAG for factual hallucination detection modeling in intelligent educational system. *2025 7th International Conference*

20. Zhong, X., Wen, S., & Kong, L. (2026). A hallucination detection framework for large language models in cybersecurity: Design and implementation. *2026 6th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, 1309–1314.
<https://doi.org/10.1109/caibda70336.2026.11621803>

Figures

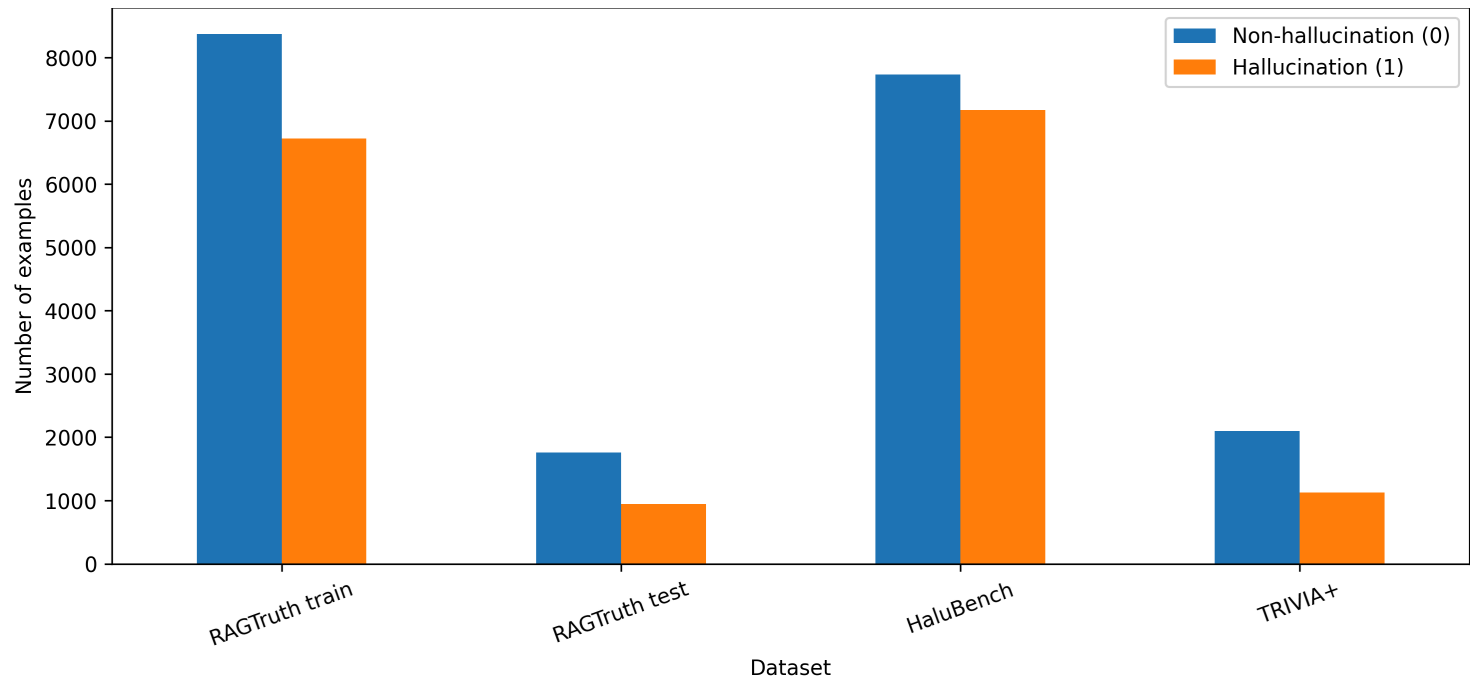


Figure 1

Class balance across datasets

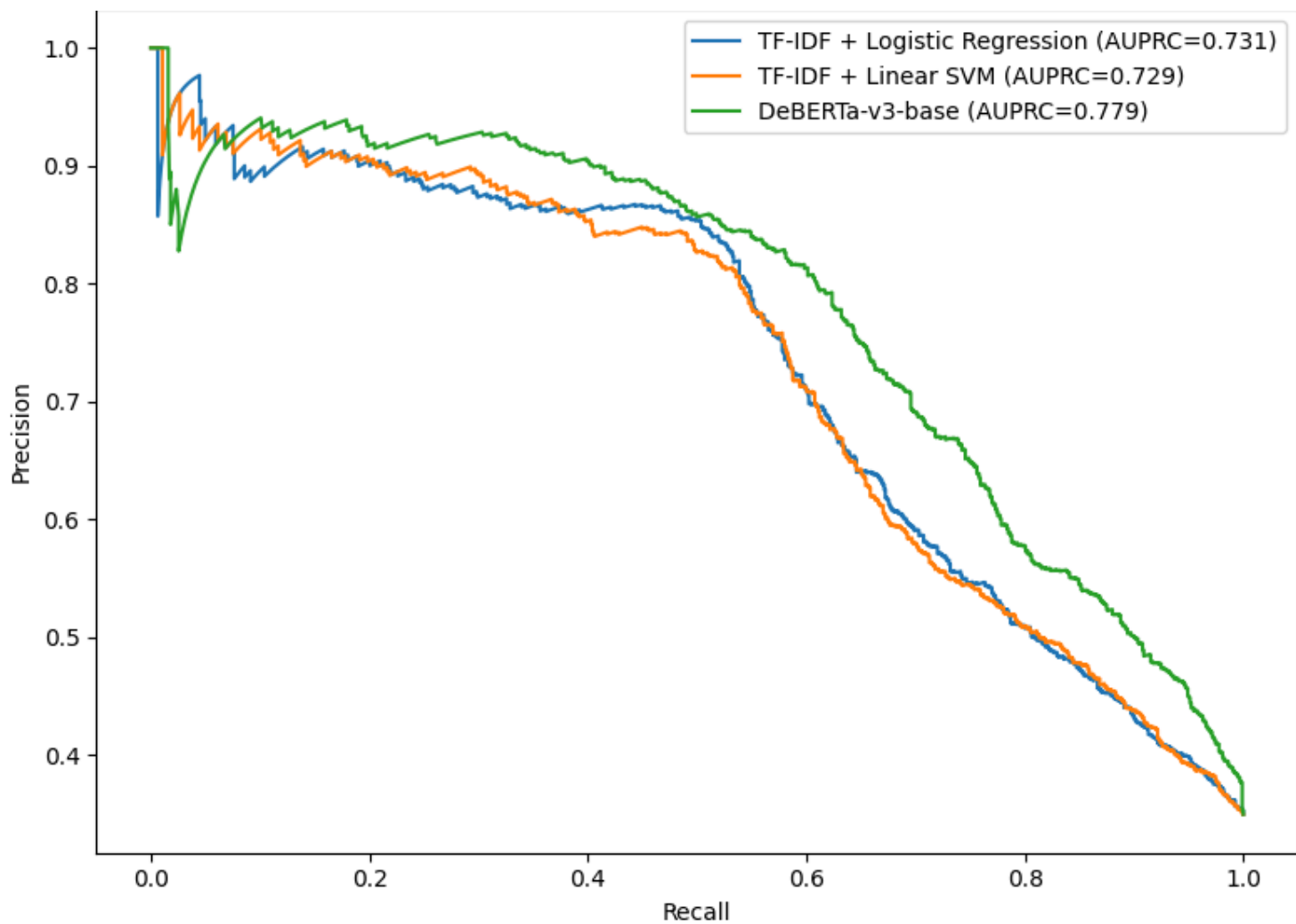


Figure 2

RQ1 Precision-Recall comparison on RAGTruth test

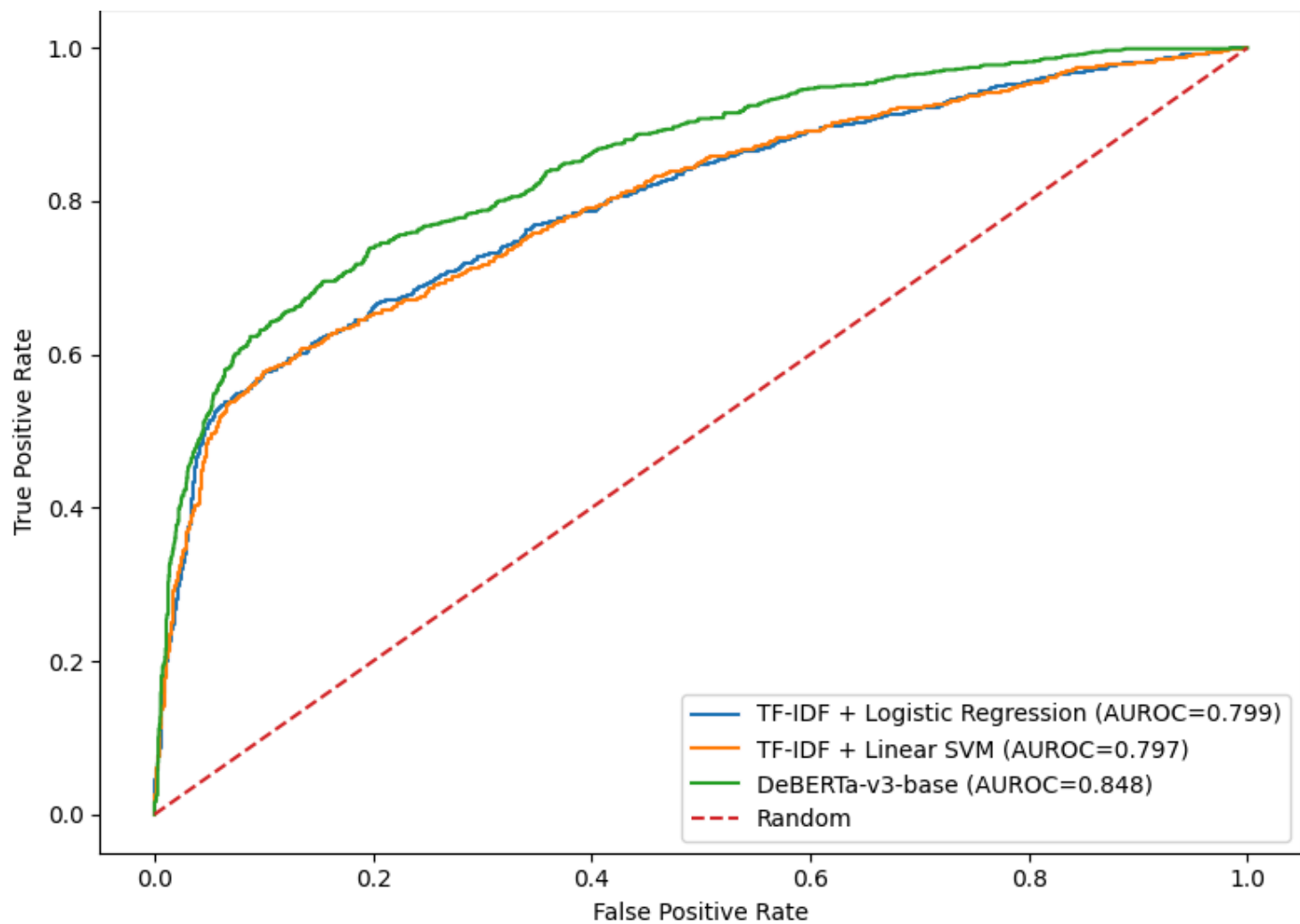


Figure 3

RQ1 ROC comparison on RAGTruth test

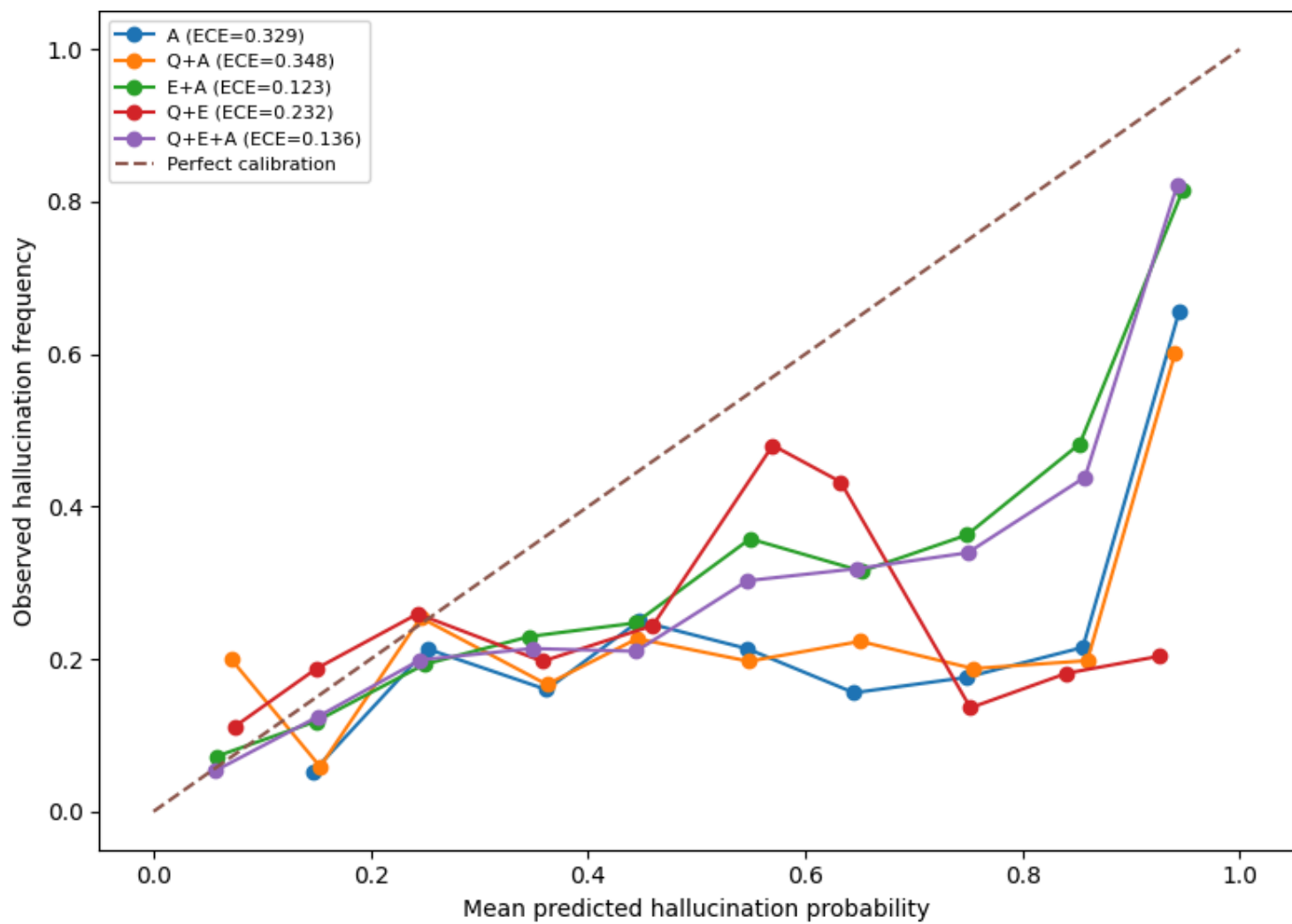


Figure 4

Calibration under input ablation