ISSN-3139-6593 (Online)

Journal of Interdisciplinary Science &Technology

Evidence-Calibrated Multi-Signal Hallucination Detection in Large Language Models

Pranav Nitin Pagare[1]*, Om Sunil Kute[2], Meet Pandurangbhai Sonvane[3]

Computer Engineering, Sandip Institute of Technology and Research Centre, Nashik[1,2]

Computer Science and Engineering, PP Savani School of Engineering[3]

pranavnpagare@gmail.com , omkute04@gmail.com, meetsonvane@gmail.com

Corresponding author-Pranav Nitin Pagare

DOI: 10.5281/zenodo.22048081

Abstract: Large language models (LLMs) can generate fluent answers that contain unsupported, contradictory, or unverifiable claims, which limits their use in knowledge-intensive and high-stakes settings. Existing detectors tend to focus on one signal, such as retrieved evidence, sampling consistency, uncertainty or internal representations. Thus they suffer differential failure when the evidence is incomplete, the model is confidently wrong or semantically similar answers differ lexically. In this paper, we propose a system Evidence-Calibrated Multi-Signal Hallucination Detection (EC-MHD) to evaluate generated text at the atomic-claim level with five complementary signals, namely evidence entailment, contradiction evidence, semantic entropy, self-consistency and confidence. These signals are fed into a calibrated fusion layer that maps them to a claim-level hallucination risk which is aggregated into a response-level decision with a specific abstention region for ambiguous circumstances. The framework is model-agnostic and can be applied to black-box LLMs with sampling and verbalized confidence, or white-box probability or hidden-state signals if available. We define a reproducible evaluation protocol spanning TruthfulQA, HaluEval, HaluEval 2.0, RAGTruth, FactBench, and process-oriented diagnostics, with accuracy, F1, AUROC, AUPRC, calibration error, span-level overlap, and latency as complementary measures. The study contributes a rigorous architecture, mathematical formulation, ablation plan, and failure-analysis procedure without fabricating empirical outcomes; numerical claims are reserved for future implementation and benchmark execution.

Keywords: Large Language Models, Hallucination Detection, Factuality, Retrieval-Augmented Verification, Semantic Entropy, Self-Consistency, Confidence Calibration, Explainable AI

1. INTRODUCTION

Large language models have gone from research demos to applications in search, writing, software engineering, analytics, teaching and domain-specific decision support. The main trust issue is not just that an answer might be wrong, but rather that a wrong answer can be given with the same fluency and confidence as a correct one. Hallucinations are described in the literature as statements that contradict a source, claims that cannot be supported by the available information, and internally inconsistent or invented features [1], [2]. This places a verification burden on users since surface quality is not a reliable proxy for factuality.

Research in benchmarking has made the problem quantifiable. TruthfulQA assesses whether a model avoids common falsehoods and misconceptions [3]. HaluEval provides a variety of hallucinated and non-hallucinated examples in general queries and task-specific contexts [4]. HaluEval 2.0 builds upon the empirical investigation of detection, causes and mitigation [11]. RAGTruth shows that retrieval-augmented generation does not solve the problem: even with external documents given, supported or conflicting spans can still exist [10]. More recent resources, such as FactBench and PROBE, move the evaluation to real-world prompts and process-level diagnostics, rather than a single final score [18], [20].

This has led to a diversification of detection methods. SelfCheckGPT uses the agreement of independently sampled generations as a black-box factuality signal [5]. FActScore decomposes long-form text into atomic facts and checks their support in a knowledge source [6]. Work on internal states suggests hidden representations can encode information about truthfulness [7], and semantic-entropy approaches estimate uncertainty by clustering outputs by meaning rather than phrasing [9]. Retrieval-based and span-level systems explicitly evaluate generated assertions against external evidence [10, 17]. These strategies complement each other but none is reliable for every operational situation. Retrieval may fail if key data is not retrieved, consistency may be high if a model keeps repeating the same misconception, confidence may be poorly calibrated, and internal-state detectors may not generalize to closed or unseen models [2], [12], [16].

This motivates a detector that views the hallucination evaluation as evidence aggregation, not a single signal classification problem. The proposed EC-MHD paradigm decomposes an answer into atomic claims and analyzes each claim via independent evidence, uncertainty, consistency and confidence channels. The channels are calibrated and scaled to comparable scales before fusion. The response is robust only in case of low risk and sufficiently informative evidence. Ambiguous cases are not forced to a binary choice but are instead sent to an abstention state.

The main contributions of this work are:

1. A multi-signal hallucination detection architecture at the claim level, which integrates retrieval evidence, semantic uncertainty, self-consistency and confidence, and applicable to both black-box and white-box LLMs.

2. A mathematical description of a calibration and fusion process that preserves high-risk assertions, as opposed to diluting them with otherwise correct material.

3. Decision output with supporting passages, contradiction evidence, uncertainty and signal contributions for each prediction that is traceable to evidence.

4. An approach to assess reproducibility, ablation, calibration, robustness and latency by using known and recent hallucination resources.

2. RELATED WORK

2.1 Hallucination Taxonomy and Evaluation

Early synthesis work demonstrated that hallucination is not a single error class. Ji et al. surveyed the unfaithfulness and unfactuality failure cases in the natural language generation tasks, while Huang et al. later organized the LLM hallucination from causes, detection, benchmarks, mitigation and knowledge boundaries [1], [2]. This distinction has operational importance: a response can be faithful to retrieved context but factually stale, or factually correct but not supported by the context provided to the model. A detector intended for deployment should therefore retain separate evidence-support and contradiction signals rather than compress all cases into a generic similarity score.

TruthfulQA [3], HaluEval [4], and HaluEval 2.0 [11] provide complementary views of factual reliability. HaluEval contains 35,000 hallucinated/normal samples across general and task-specific settings, while RAGTruth contains nearly 18,000 naturally generated RAG responses with case- and word-level annotations [10]. FactBench uses 1,000 prompts across 150 topics derived from real-world model usage and explicitly recognizes Supported, Unsupported, and Undecidable units [18]. PROBE further decomposes the detection pipeline into claim decomposition, evidence finding, evidence evaluation, and hallucination localization across 12,000 test cases [20]. These resources support the design choice in EC-MHD to expose intermediate verification stages rather than return only a final label.

2.2 Sampling, Uncertainty, and Internal Signals

SelfCheckGPT demonstrated that factual and non-factual content can often be separated by comparing a target output with multiple samples from the same black-box model [5]. The method is attractive because it does not require an external database or model internals. However, sampling increases inference cost, and agreement is not equivalent to truth: a model can consistently reproduce the same false association. Mündler et al. similarly showed that self-contradiction is a useful diagnostic and can be detected without external retrieval, but self-consistency and factual correctness remain distinct objectives [13].

Another useful, if flawed, signal is uncertainty. Xiong et al. found that verbalized confidence is often overconfident and that no elicitation strategy is uniformly best across tasks [12]. Farquhar et al. proposed semantic entropy where generations are clustered based on semantic equivalence before entropy calculation, which reduces sensitivity to superficial variations in wording [9]. White-box approaches provide more options: Azaria and Mitchell showed that internal activations are useful for truthfulness classifiers [7], and PRISM later used prompt-guided internal states to improve cross-domain detection [16]. These studies motivate a modular confidence interface in EC-MHD: when available, white-box probability or representation features can be used but the detector must remain functional with black-box access only.

2.3 Evidence-Grounded Detection and Mitigation

Retrieval-augmented generation proposed a general mechanism for conditioning the generation on non-parametric evidence [8]. The key shift for hallucination detection is to go from retrieving information to evaluating whether specific generated claims are supported by that information. RAGTruth provides a benchmark for this setting [10]. REFIND uses retrieved documents directly for detection of span-level hallucination in multiple languages [17]. A related principle is operationalized by FActScore that breaks down long text into atomic facts and measures the proportion of them supported by a reliable knowledge source [6].

Grounding evidence also interacts with the decoding and model-side interventions. DoLa improves the factuality by comparing distributions from different layers without retrieval [14], while HaloScope learns a truthfulness classifier from unlabeled generations in the wild [15]. FactSelfCheck pushes black-box sampling to fact-level granularity and reports that fine-grained localization improves correction comparing with sentence-level checking [21]. These developments indicate that the best design direction is not to choose one family exclusively but to mix complementary channels while preserving intermediate outputs interpretable.

2.4 Research Gap

The literature identifies four persistent gaps. First, we show that many detectors are designed for a single access regime: retrieval, sampling, or internal states, and degrade when that resource is not available. Second, the fusion of scores is often implicit, various signals are combined without rigorous calibration . Third, passage-level decisions can mask a single high risk claim within a correct answer. Fourth, discrimination metrics, calibration, abstention behavior, retrieval failure, and end-to-end delay are not usually reported in evaluation. These limitations motivate the EC-MHD: Atomic claims are the unit of analysis; Signal reliability is calibrated on held-out data; Response risk is tail-sensitive; and The evaluation protocol treats uncertainty and computational cost as first-class outcomes.

Table 1. Representative Literature and the Design Implication for EC-MHD

Study	Primary signal / resource	Granularity	Key contribution	Limitation relevant to this work
TruthfulQA [3]	Truthfulness benchmark	Answer	Tests model tendency to reproduce common falsehoods	Benchmark only; does not localize evidence support
SelfCheckGPT [5]	Sampling consistency	Sentence	Black-box detection without external knowledge	Multiple generations add cost; consensus can still be wrong
FActScore [6]	Atomic-fact evidence support	Fact	Fine-grained factual precision against knowledge sources	Quality depends on decomposition and evidence retrieval
Semantic Entropy [9]	Meaning-level uncertainty	Answer / passage	Separates semantic variation from lexical variation	Targets uncertainty/confabulation rather than every factual error
RAGTruth [10]	RAG hallucination corpus	Case + word	Natural RAG responses with fine-grained annotations	Domain/task coverage is finite
PRISM [16]	Prompt-guided internal states	Response	Improves supervised detector transfer across domains	Requires access to internal representations

FactBench [18]	Dynamic real-world factuality	Content unit	Real-world prompts and supported/unsupported/undecidable labels	Requires evidence retrieval and periodic refresh
PROBE [20]	Process-level benchmark	Pipeline stage	Separates decomposition, retrieval, evaluation, and localization	Diagnostic benchmark; not itself a unified deployment detector
FactSelfCheck [21]	Fact-level black-box sampling	Fact	Fine-grained sampling-based detection and correction	Sampling remains computationally expensive

3. PROPOSED EC-MHD FRAMEWORK

3.1 Problem Formulation

Let q be a user query, R be the response generated by a target LLM. The claim decomposition function g(·) maps R to a set of atomic claims C = {c1, c2, ..., cn}. Each claim is assessed against retrieved evidence and model-derived uncertainty signals. The goal is to estimate a hallucination probability h i ∈ [0,1] for each claim and a response level risk H_R , while maintaining the trace of evidence that explains the decision. The framework defines three outcomes: Reliable, Hallucinated and Abstain. Abstention plays a key role when there is not enough evidence for retrieval or calibrated risk falls in a deliberately uncertain range.

3.2 Claim Decomposition and Evidence Retrieval

Atomic decomposition reduces the masking effect of mixed-quality passages. A long response may contain many correct statements and one fabricated number, date, attribution, or causal claim; passage-level similarity can overlook that local error. EC-MHD therefore segments the response into minimally verifiable propositions. Non-verifiable discourse markers, subjective preferences, and purely stylistic text are marked as non-factual units and excluded from factual scoring. This mirrors the motivation behind atomic-fact evaluation [6] and process-level diagnostics [20].

For each factual claim c_i, a retrieval module returns the top-k candidate passages D_i={d_i1,...,d_ik}. Retrieval quality is tracked separately from factual support. This prevents the system from treating 'no evidence found' as equivalent to 'claim disproved'. The retrieval-quality variable q_i summarizes document relevance and coverage; if q_i is below a minimum threshold, the final decision is biased toward Abstain rather than Hallucinated.

$$s_e(c_i) = max_{d\in D_i} P(entail|d, c_i), s_c(c_i) = max_{d\in D_i} P(contradict|d, c_i) \quad (1)$$

Equation (1) separates positive support from explicit contradiction. Neutral or irrelevant evidence should not be forced into either class. This is especially important for open-domain questions where absence of evidence may reflect retrieval failure rather than falsity.

3.3 Semantic Entropy and Self-Consistency

The target LLM is sampled m times under controlled stochastic decoding. Sampled answers are clustered by semantic equivalence so that paraphrases contribute to the same meaning cluster. If p_j is the empirical probability mass of semantic cluster j, semantic entropy is computed as:

$$H_sem(q) = -\Sigma_j p_j log(p_j), \hat{H}_sem = H_sem/log(K) \quad (2)$$

where K is the number of observed semantic clusters and Ĥ_sem ∈ [0,1] is the normalized semantic uncertainty. This follows the central insight that uncertainty should be measured over meanings rather than exact strings [9]. EC-MHD additionally computes claim consistency against the sampled answers. Let I(c_i, R^(a), R^(b)) be an agreement indicator or semantic entailment score for the claim across two samples; then:

$$S_i = 2/[m(m-1)]\Sigma_{a<b} I(c_i, R^{(a)}, R^{(b)}) \quad (3)$$

High consistency lowers risk only when other channels do not contradict the claim. This asymmetric treatment prevents repeated misinformation from being accepted merely because the model is stable.

3.4 Confidence and Calibration

We use confidence as an auxiliary signal and not as ground-truth. For white box models, token probabilities or hidden state classifiers [7], [16] can be used. In black-box systems, EC-MHD depends on verbalized confidence, sampling frequency or a light-weight external confidence estimator. Since the raw confidence scales differ for different models and tasks, the signal is calibrated on a held-out validation set using temperature scaling, isotonic regression or another monotonic calibration map . Let u_i be the raw confidence-derived uncertainty and T(·) the learned calibration map:

$$U_i = T(u_i), 0 \le U_i \le 1 \quad (4)$$

The framework evaluates Expected Calibration Error (ECE) and reliability diagrams in addition to accuracy. A detector that ranks hallucinations well but produces unreliable probabilities is unsuitable for risk-thresholded deployment.

3.5 Calibrated Multi-Signal Fusion

The final claim risk combines evidence deficit, contradiction, semantic uncertainty, inconsistency, confidence uncertainty, and retrieval quality. Rather than fixing equal weights, EC-MHD learns non-negative coefficients on validation data with regularization and then calibrates the fused output. One practical formulation is:

$$h_i = \sigma\left(\beta_0 + \beta_1(1-s_e) + \beta_2 s_c + \beta_3 \hat{H}_s em + \beta_4(1-S_i) + \beta_5 U_i + \beta_6(1-q_i)\right) \quad (5)$$

where σ(·) is the logistic function. The signs encode an interpretable prior: lack of entailment, contradiction, uncertainty, inconsistency, confidence uncertainty, and weak retrieval increase hallucination risk. These constraints can be relaxed only if empirical validation justifies a different relationship.

Response-level aggregation must preserve isolated high-risk claims. A simple arithmetic mean can dilute one serious hallucination inside a long answer, so EC-MHD combines weighted mean risk with the maximum claim risk:

$$H_R = \alpha max_i(h_i) + (1-\alpha)\Sigma_i \omega_i h_i, \Sigma_i \omega_i = 1 \quad (6)$$

where α controls sensitivity to local failures and ω_i can reflect claim importance or verifiability. Decision thresholds τ_low and τ_high define the three-way output: Reliable if H_R ≤ τ_low, Hallucinated if H_R ≥ τ_high, and Abstain otherwise. Thresholds are selected on validation data according to the target false-negative cost rather than tuned on the test set.

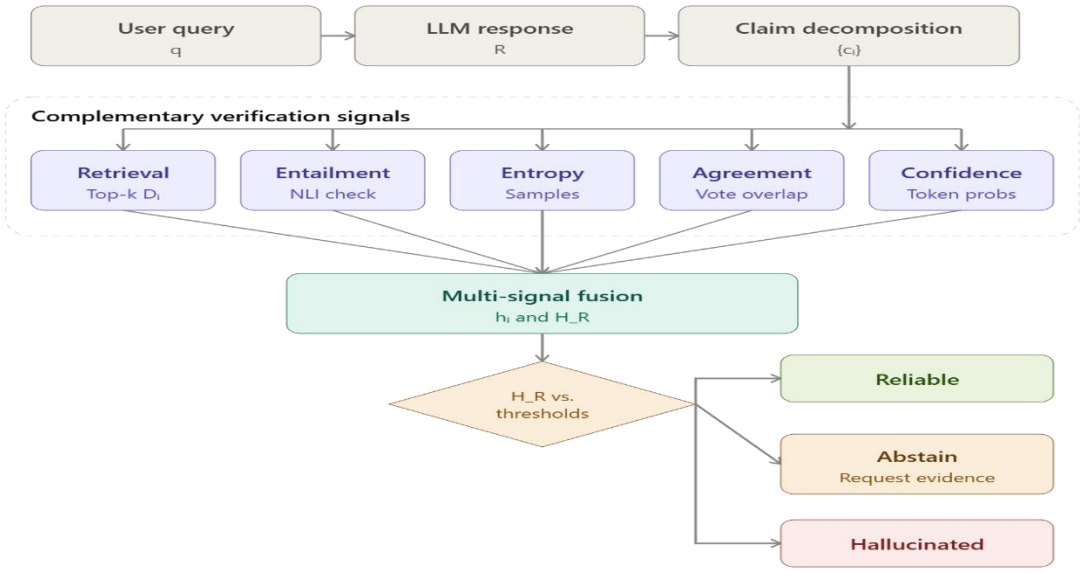

Figure 1. Evidence-Calibrated Multi-Signal Hallucination Detection (EC-MHD) architecture. Each generated response is decomposed into atomic claims; complementary evidence, uncertainty, consistency, and confidence channels are fused after calibration before a three-way decision

3.6 Algorithmic Procedure

The detector is described without assuming a particular LLM vendor, retrieval engine or entailment model in Algorithm 1:

1. get query q and get target response R from the specified LLM.

2. Decompose R into atomic verifiable claims C and label non-factual units .

3. For each claim, retrieve top-k candidate evidence passages and estimate retrieval quality q_i.

4. For each claim compute entailment support s e and contradiction support s_c.

5. Generate m controlled samples, Cluster the semantically identical results, and compute normalized semantic entropy.

6. Estimate the claims' self-consistency across samples, and get the confidence signal available.

7. For individual channels, calibrate on held-out data and compute h_i, the fused claim risk.

8. H_R = Aggregate claim risks; compare with τ low and τ high to return Reliable, Hallucinated or Abstain.

9. Provide label with evidence at claim level, risk values and signal contributions for auditability.

4. EXPERIMENTAL DESIGN

4.1 Benchmark Suite

Rather than a single benchmark, the proposed system should be tested on datasets highlighting various failure mechanisms. TruthfulQA measures the robustness to repeated false statements [3] . HaluEval offers both general and task-specific hallucination examples [4]. More recently, factuality analysis support is provided by HaluEval 2.0 [11]. We evaluate hallucinations in retrieval-augmented generation with RAGTruth providing fine-grained annotations [10]. FactBench adds real world prompts and an Undecidable category [18]. PROBE is used for diagnosis at component level, evidence collection, evidence evaluation and localization in the decomposition [20]. MedHallu can be used as a domain-specific stress test, where small factual mistakes carry bigger practical risk [19].

Table 2. Proposed Benchmark and Evaluation Mapping

Benchmark	Verified scope / scale	Primary evaluation role	Recommended split use
TruthfulQA [3]	817 questions across 38 categories	Open-domain truthfulness and calibration	Official/test protocol; no training leakage
HaluEval [4]	35,000 hallucinated/normal samples	General, QA, dialogue, summarization detection	Use official partitions; report task-wise metrics
HaluEval 2.0 [11]	Updated factuality hallucination benchmark	Detection robustness and failure analysis	Held-out test; compare with HaluEval
RAGTruth [10]	Nearly 18,000 RAG responses; case + word annotations	Retrieval-grounded and span-level detection	Case-level F1 + span overlap/IoU
FactBench [18]	1,000 prompts across 150 topics	In-the-wild factuality and undecidable cases	Use benchmark tiers; report refusal/abstention
PROBE [20]	12,000 process-level cases	Diagnose decomposition, evidence finding/evaluation, localization	Component-level evaluation
MedHallu [19]	10,000 medical QA pairs	Domain-specific stress test	Report easy/medium/hard tiers separately

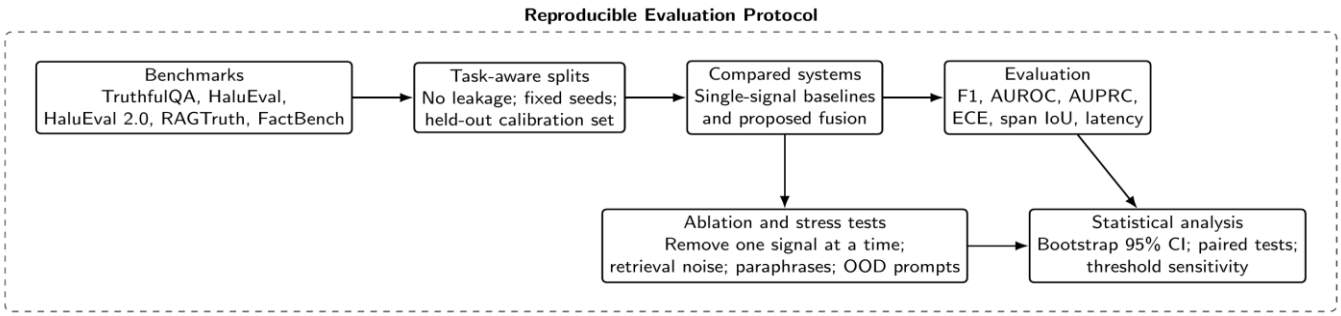

Figure 2. Reproducible evaluation pipeline. Benchmarks, split control, baseline comparison, calibration, ablation, stress testing, and statistical analysis are treated as one protocol rather than independent afterthoughts.

4.2 Baselines and Ablation

The comparison set should include at least one representative from each of the major signal families: lexical or embedding similarity, retrieval-only entailment, SelfCheckGPT-style sampling [5], semantic entropy [9], confidence-only detection [12], and a claim-evidence factuality baseline such as FActScore [6]. When model internals are available, an internal-state detector can be added [7], [16]. The main comparison should be performed on the same data partitions and target outputs, so that differences can be attributed to detection quality rather than different generation conditions.

Ablation removes one EC-MHD channel at a time: evidence entailment, contradiction, semantic entropy, self-consistency, confidence, retrieval quality gating, and maximum risk aggregation. A second ablation compares fixed equal weights with learned calibrated weights.

A third one compares binary classification-versus the three-way decision Reliable/Hallucinated/Abstain. The goal is to find the strongest variant, but also to find which component contributes under which failure condition.

4.3 Metrics and Statistical Analysis

Accuracy alone is not enough as hallucination datasets can be imbalance and deployment costs are asymmetric. The main classification metrics are Precision, Recall, F1, AUROC and AUPRC. RAGTruth also supports span-level overlap or Intersection-over-Union The Brier score and the ECE are used to evaluate calibration. Selective prediction is evaluated with coverage-risk curves: as the model abstains on uncertain cases, residual error among answered cases should decrease. Efficiency is reported using mean and percentile latency, number of LLM calls per query, retrieval calls, and token usage.

$$Precision = TP/(TP + FP),$$

$$Recall = TP/(TP + FN), F1 = 2PR/(P + R) \quad (7)$$

$$ECE = \Sigma_b(|B_b|/N)|acc(B_b) - conf(B_b)| \quad (8)$$

Bootstrap resampling at the query level will be used to compute 95% confidence intervals for each principal metric. For comparisons between systems on the same examples, paired bootstrap or an appropriate paired test should be used. It is important to select thresholds, fusion weights, calibration maps and stopping criteria only on the training/validation data. Results should also be stratified by task, response length, retrieval quality and hallucination type so that failure modes are not hidden within aggregate scores.

4.4 Robustness and Reproducibility

Robustness tests should vary only one factor at a time. Retrieval stress tests can remove top ranked passage, add irrelevant documents or decrease top-k. Stress tests can paraphrase the query promptly, without loss of meaning. Sampling stress tests vary number of generations and temperature. Cross-domain tests train calibration parameters on a subset, and evaluate on the other subset. Reproducibility requires versioned datasets, fixed seeds where stochasticity is controllable, logged model identifiers, prompt templates, retrieval configuration and a machine readable record of each claim/evidence decision.

5. DISCUSSION

5.1 Why Multi-Signal Detection Is Necessary

No single channel should be taken as a factual oracle. Retrieval offers external grounding, but may fail to retrieve the correct evidence. Self-consistency is cheap to interpret but expensive to sample, and can reinforce a stable misconception. Semantic entropy shows epistemic uncertainty, but does not show falsity.

Confidence may help to triage, but it is often miscalibrated [12]. Internal representations can be informative but are not available for proprietary models [7], [16]. EC-MHD treats them as semi-independent observations and uses calibration to learn their reliability [22].

5.2 Explainability and Human Review

The explanation of the framework is not free-form but evidence-centric. For each flagged claim, the system can show the retrieved passages, entailment/contradiction values, semantic uncertainty, consistency score, confidence uncertainty and the contribution of each channel to h_i. This allows a false positive diagnosis: the reviewer can determine whether the error came from claim decomposition, retrieval, evidence assessment or the fusion threshold. The process structure complies with the diagnostic motivation of PROBE [20] and recent fine-grained detectors [17], [21-24].

5.3 Limitations

The proposed framework does not remove hallucination. Its effectiveness is constrained by the quality and freshness of evidence sources, the accuracy of claim decomposition and entailment models, and the availability of sufficient samples for uncertainty estimation. Some statements are subjectively ambiguous, temporally dynamic, or not practically verifiable; in such cases, forcing a binary truth label would be a kind of error itself. Thus Abstain is an outcome of the method rather than a failure to decide. One limitation is the computational cost, as retrieval and repeated sampling can increase latency. Deployment needs to support adaptive computation, e.g. only running expensive semantic-entropy checks when cheaper evidence channels disagree.

Another limitation is that this paper does not provide benchmark results, although it defines a methodological framework and a reproducible experimental protocol. The only time numerical performance should be reported is in the case of the fixed protocol. The difference is important for research integrity because plausible looking synthetic results would detract (not add) to the contribution.

6. CONCLUSION

This paper proposed EC-MHD, an evidence-calibrated multi-signal framework for hallucination detection in large language models. The design includes atomic-claim decomposition, support for external evidence and contradiction, semantic entropy and self-consistency, and calibrated confidence. Claim-level risks are aggregated into a tail-sensitive response score and a three-way decision rule permitting abstention when evidence is insufficient or conflicting signals are observed. We evaluate the paper on benchmark diversity, calibration, span-level localization, ablation, robustness, latency, and statistical uncertainty. The next step is the implementation and benchmark execution under the defined protocol. Next, the decision of threshold for application-specific risk profiles. Our main study hypothesis is that trustworthy hallucination detection needs calibrated agreement between evidence, uncertainty, and consistency signals with transparent failure tracing, and not a universal score.

REFERENCES

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, Art. 248, 2023, doi: 10.1145/3571730.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Transactions on Information Systems, vol. 43, no. 2, pp. 1–55, 2025, doi: 10.1145/3703155.

[3] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 2022, pp. 3214–3252, doi: 10.18653/v1/2022.acl-long.229.

[4] 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, Singapore, 2023, pp. 6449–6464, doi: 10.18653/v1/2023.emnlp-main.397.

[5] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in Proc. EMNLP, Singapore, 2023, pp. 9004–9017, doi: 10.18653/v1/2023.emnlp-main.557.

[6] 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, Singapore, 2023, pp. 12076–12100, doi: 10.18653/v1/2023.emnlp-main.741.

[7] A. Azaria and T. Mitchell, "The Internal State of an LLM Knows When It's Lying," in Findings of EMNLP, Singapore, 2023, pp. 967–976, doi: 10.18653/v1/2023.findings-emnlp.68.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474.

[9] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting Hallucinations in Large Language Models Using Semantic Entropy," Nature, vol. 630, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.

[10] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang, "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in Proc. 62nd ACL, Bangkok, Thailand, 2024, pp. 10862–10878, doi: 10.18653/v1/2024.acl-long.585.

[11] J. Li, J. Chen, R. Ren, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models," in Proc. 62nd ACL, Bangkok, Thailand, 2024, pp. 10879–10899, doi: 10.18653/v1/2024.acl-long.586.

[12] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs," in Proc. International Conference on Learning Representations (ICLR), 2024.

[13] N. Mündler, J. He, S. Jenko, and M. Vechev, "Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation," in Proc. International Conference on Learning Representations (ICLR), 2024.

[14] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He, "DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models," in Proc. International Conference on Learning Representations (ICLR), 2024.

[15] X. Du, C. Xiao, and Y. Li, "HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection," in Advances in Neural Information Processing Systems, vol. 37, 2024.

[16] F. Zhang, P. Yu, B. Yi, B. Zhang, T. Li, and Z. Liu, "Prompt-Guided Internal States for Hallucination Detection of Large Language Models," in Proc. 63rd ACL, Vienna, Austria, 2025, pp. 21806–21818, doi: 10.18653/v1/2025.acl-long.1058.

[17] D. Lee and H. Yu, "REFIND at SemEval-2025 Task 3: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models," in Proc. 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria, 2025, pp. 7–15.

[18] F. Fatahi Bayat, L. Zhang, S. Munir, and L. Wang, "FactBench: A Dynamic Benchmark for In-the-Wild Language Model Factuality Evaluation," in Proc. 63rd ACL, Vienna, Austria, 2025, pp. 33090–33110, doi: 10.18653/v1/2025.acl-long.1587.

[19] S. Pandit, J. Xu, J. Hong, Z. Wang, T. Chen, K. Xu, and Y. Ding, "MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models," in Proc. EMNLP, Suzhou, China, 2025, pp. 2858–2873, doi: 10.18653/v1/2025.emnlp-main.143.

[20] Y. Zhang, P. Belcak, S. Diao, Y. Fu, S. Ghosh, M. Mardani, E. M. P. Long, B. Yu, and P. Molchanov, "PROBE: PROcess-Based BEnchmark for Hallucination Detection," in Findings of ACL, San Diego, CA, USA, 2026, pp. 42303–42320, doi: 10.18653/v1/2026.findings-acl.2099.

[21] A. Sawczyn, J. Binkowski, D. Janiak, B. Gabrys, and T. J. Kajdanowicz, "FactSelfCheck: Fact-Level Black-Box Hallucination Detection for LLMs," in Findings of EACL, Rabat, Morocco, 2026, pp. 5603–5621, doi: 10.18653/v1/2026.findings-eacl.296.

[22] Budhewar AS, Patil BK, Tharayil AS, Bhosale S, Suthadevan S, Patel N, et al. Federated AI-Driven Digital Twins in the Healthcare Metaverse: Architectures, Privacy, and Clinical Intelligence. In: Ben Othman S, editor. The Convergence of the Metaverse, AI, and Federated Learning in Healthcare Ecosystems. IGI Global Scientific Publishing; 2026. p. 217–250. Available from: https://doi.org/10.4018/979-8-3373-8889-2.ch008

[23] Somwanshi SK, Tharayil AS, Budhewar AS, Velu G, Thamizanban D, Gayathri KC, et al. Beyond Performance Benchmarks: A Safety-Critical Evaluation Paradigm for Reliability and Trust in Clinical Language Models. In: Ortiz-Rodriguez F, Kansal M, Kumar Dhanaraj R, Khan F, editors. Medical LLMs for Clinical Safety Assessment. IGI Global Scientific Publishing; 2026. p. 173–200. Available from: https://doi.org/10.4018/979-8-3373-7862-6.ch007

[24] Veernapu KK, Patil BK, Tharayil AS, Sindhura C, Andy A, Budhewar A, et al. Real-time AI in Clinical Decision Support: Bridging Research and Practice. In: Andy A, Tharayil A, Budhewar A, editors. Revolutionizing Drug Research and Personalized Medicine Through AI and Machine Learning. IGI Global Scientific Publishing; 2026. p. 431–452. Available from: https://doi.org/10.4018/979-8-3373-6400-1.ch015

