IJMRT: Volume (8), Issue 11, 2026, Pages 1-13
International Journal of Multidisciplinary Research Transactions
(A Peer Reviewed Journal)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN(Online):2663-4007
|Regular Article |Open Access
Aravandi Publications

Improving the Reliability of Large Language Models: A Hypothesis-Driven Statistical Evaluation of Hallucination Detection and Mitigation Techniques

Mihir S. Bhogal[1*], Jigyasa Gupta[2], Aayush Pratap Singh[3], Nitish Kumar Maurya[4]

[1,2,3,4] Department of CS (AI & ML), Shri Ramswaroop Memorial College of Engg & Management, ADATU, Lucknow, India.

*Corresponding author(s):
Mihir S. Bhogal, Department of CS (AI & ML), Shri Ramswaroop Memorial College of Engg & Management, ADATU, Lucknow, India.
Email: mihir.bhogal@gmail.com

DoI: https://doi.org/10.5281/zenodo.24110351

Citation:

Mihir S. Bhogal, Jigyasa Gupta, Aayush Pratap Singh, Nitish Kumar Maurya, (2026). Improving the Reliability of Large Language Models: A Hypothesis-Driven Statistical Evaluation of Hallucination Detection and Mitigation Techniques, International Journal of Multidisciplinary Research Transactions, *8*(11), 1–13. https://doi.org/10.5281/zenodo.24110351

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Accepted: 5 November 2026 Available online: 10 November 2026

Abstract

Large Language Models (LLMs) such as GPT-4, LLaMA, and Claude achieve remarkable fluency, yet their tendency to hallucinate — producing plausible but factually incorrect content — remains a central obstacle to reliable deployment in high-stakes domains. This paper organizes a comparative evaluation of six hallucination detection techniques (SelfCheckGPT, FActScore, Chain-of-Verification, internal-state probing, semantic entropy, and Retrieval-Augmented Generation) and six mitigation techniques around two falsifiable statistical hypotheses: that mean detection accuracy (AUROC) differs across technique families (H1), and that mean mitigation effectiveness differs across technique families (H2). A pilot dataset (n=12 per technique) built from benchmark-aggregated literature values was analyzed using one-way ANOVA, Levene's test, Shapiro–Wilk normality testing, Tukey's HSD post-hoc test, and eta-squared effect size. Both omnibus ANOVAs were significant (detection: $F(5,66)=43.99$, $p<0.001$, $\eta^2=0.77$; mitigation: $F(5,66)=55.24$, $p<0.001$, $\eta^2=0.81$), rejecting both nulls. Tukey HSD resolved detection into three statistically distinct tiers led by RAG Grounding Consistency and Semantic Entropy, and resolved mitigation into a tiered structure led by RAG alone. These results motivate a modular ensemble pipeline combining

Copyrights@IJMRT

lightweight detection, retrieval grounding, and selective verification as the most defensible path toward reliable LLM systems.

Keywords: Large Language Models, Hallucination Detection, Hallucination Mitigation, Retrieval-Augmented Generation, Chain-of-Verification, Hypothesis Testing, ANOVA, Reliability, Trustworthy AI.

1. Introduction

The rise of LLMs in search, healthcare, legal, and software applications has raised concern over hallucination: fluent, confident text that is factually wrong or unverifiable. Because such output is linguistically well-formed, it is difficult to detect and dangerous where users do not independently verify claims. Hallucination is increasingly viewed as an emergent property of training LLMs as next-token predictors on imperfect data and tuned for fluency rather than calibrated truthfulness; Xu et al. [4] argue it may be an inherent, unavoidable property of any computable bounded-knowledge language model, shifting the research question from elimination to detection and containment.

This paper presents a comparative, hypothesis-driven statistical evaluation of major hallucination detection and mitigation technique families. Rather than proposing a new algorithm, it contributes: (1) a structured comparison across accuracy, cost, and complexity; (2) two formally stated statistical hypotheses; (3) a documented statistical methodology; and (4) a modular pipeline informed by the resulting rankings. The rising cost of undetected hallucinations in decision-support pipelines — false citations, drug-interaction errors, invented figures — motivates the need for statistically defensible, not merely descriptive, evidence of which techniques justify their added cost.

1.1. Objectives

- Review and categorize current hallucination detection and mitigation techniques for LLMs.
- Formally test whether detection accuracy and mitigation effectiveness differ significantly across technique families.
- Identify, via post-hoc testing, which techniques are statistically distinguishable from which others.
- Propose a modular, ensemble-based pipeline informed by the rankings, and outline a future research agenda.

2. Literature Review

Ji et al. [1] provide one of the first comprehensive surveys of hallucination in natural language generation, distinguishing intrinsic (source-contradicting) from extrinsic (source-unverifiable) hallucination — a taxonomy adopted throughout this paper. **Zhang et al.** [2] trace hallucination to data-, training-, and inference-related causes, and caution that Retrieval-Augmented Generation does not eliminate hallucination since retrieval failures can introduce new errors. **Huang et al.** [3] catalog benchmarks across QA, summarization, dialogue, and code generation, noting that fragmented metrics make cross-technique comparison difficult — the gap this paper targets. **Xu et al.** [4] relate hallucination to undecidability results in computation theory, reframing the goal from elimination toward detection and containment.

Within detection, **SelfCheckGPT** [5] samples multiple stochastic generations and measures mutual consistency, at the cost of expensive repeated sampling — a decoding-time idea that echoes self-consistency sampling for reasoning tasks [29]. **FActScore** [6] decomposes generations into atomic facts verified against a knowledge base, trading completeness of that base for interpretability. **Azaria and Mitchell** [7] probe hidden activations to flag falsehoods the model has implicitly "recognized," building on evidence that LLMs are often reasonably well calibrated about what they do and do not know [28], and **Orgad et al.** [8] show internal representations often encode more accurate information than the surface text expresses. Mündler et al. [9] detect self-contradictory hallucinations via purely internal consistency checks. Semantic-entropy methods cluster generations by meaning and compute entropy over the resulting clusters [26], a formulation later shown to detect confabulations in state-of-the-art LLMs without task-specific supervision [25]; the approach offers robustness to paraphrase and emerges as one of the top-performing techniques in this paper's comparison.

On mitigation, **Shuster et al.** [10] show retrieval augmentation reduces hallucination relative to purely parametric baselines, extending the original Retrieval-Augmented Generation architecture [24], though noisy or poorly ranked retrieval can itself introduce errors — a limitation sharpened by evidence that LLMs use relevant information less reliably when it is buried in the middle of a long retrieved context [30]. **Dhuliawala et al.** [11] propose Chain-of-Verification (CoVe), drafting an answer, generating independent verification questions, and revising based on discrepancies — effective but latency-costly. Pan et al. [12] find self-correction driven by a model's own critique is unreliable, while correction using external or independently sampled evidence is markedly more effective, consistent with CoVe's design.

Zhang, Yu and Feng [13] propose TruthX, editing a learned "truthful space" in internal representations at inference time. **Tonmoy et al.** [14] survey over thirty mitigation techniques and conclude hybrid pipelines combining complementary methods are already the practical norm — a conclusion this paper's proposed ensemble operationalizes with statistical justification.

Beyond these core families, **de Almeida da Silva et al.** [15] find domain-restricted retrieval reduces error in educational applications; **Zhang et al.** [16] show LLMs invent plausible but non-existent APIs in code generation, a failure mode catchable via compilation; Perković et al. [17] highlight that organizational risk tolerance and auditability shape which techniques are practical in enterprise deployment; and **Yehuda et al.** [18] propose InterrogateLLM, a zero-resource detection method based on question-reconstruction consistency. Much of this evaluation ultimately rests on truthfulness benchmarks such as **TruthfulQA** [27], which shows models can confidently reproduce common misconceptions rather than genuinely known facts. Collectively, this literature establishes an agreed taxonomy [1], several mechanistically distinct technique families [5]–[13], [24]–[26], [28]–[29], and validation across domains [15]–[18], but no study places these families side by side under one evaluation lens and formally tests, via inferential statistics, whether observed differences are significant — the gap Sections 3–6 address.

3. Research Methodology

This work follows a comparative, literature-synthesis evaluation supported by a proposed architecture and formal hypothesis testing, rather than training new models. Reported performance figures for six techniques (Table 1) were aggregated and normalized onto common evaluation dimensions — detection accuracy (AUROC, F1), mitigation effectiveness (% hallucination-rate reduction vs. an ungrounded baseline), and computational overhead (latency relative to a single-pass baseline) — then subjected to inferential testing (Sections 4–6). Techniques were selected for citation frequency and adoption (2022–2025), availability of quantitative benchmarks, and coverage of three intervention points: pre-generation (retrieval), in-generation (decoding constraints), and post-generation (detection/verification). Where studies overlapped, mean reported values were used to reduce single-study bias.

Table.1. Comparative Summary of Hallucination Detection Techniques

Technique	Category	Core Mechanism	AUROC/ F1	Latency	Key Limitation

SelfCheckGPT	Detection	Sampling-based consistency	0.75/0.71	~4.8×	Costly; weak on deterministic errors
FActScore	Detection	Atomic facts vs. KB	0.73/0.69	~3.2×	Needs reliable KB; slow
Chain-of-Verification	Both	Self-generated verification Qs	0.81/0.78	~5.5×	High latency; Qs can be flawed
Internal-State Probing	Detection	Probes on hidden activations	0.79/0.74	~1.3×	White-box only; weak generalization
Semantic Entropy	Detection	Clusters outputs by meaning	0.83/0.80	~2.1×	Threshold-sensitive; costly sampling
RAG Grounding Consistency	Both	Cross-checks retrieved evidence	0.87/0.84	~2.6×	Inherits retriever errors; stale corpora

Beyond this synthesis, we propose a modular pipeline (Figure 1) routing each response through parallel detection, grounding, and verification layers feeding a weighted risk aggregator; responses below threshold pass directly, while flagged responses undergo regeneration, retrieval-augmented correction, or abstention. Table 1's values are illustrative, synthesizing reported literature rather than one controlled test set; Section 8 outlines the standardized benchmarking needed for full resolution.

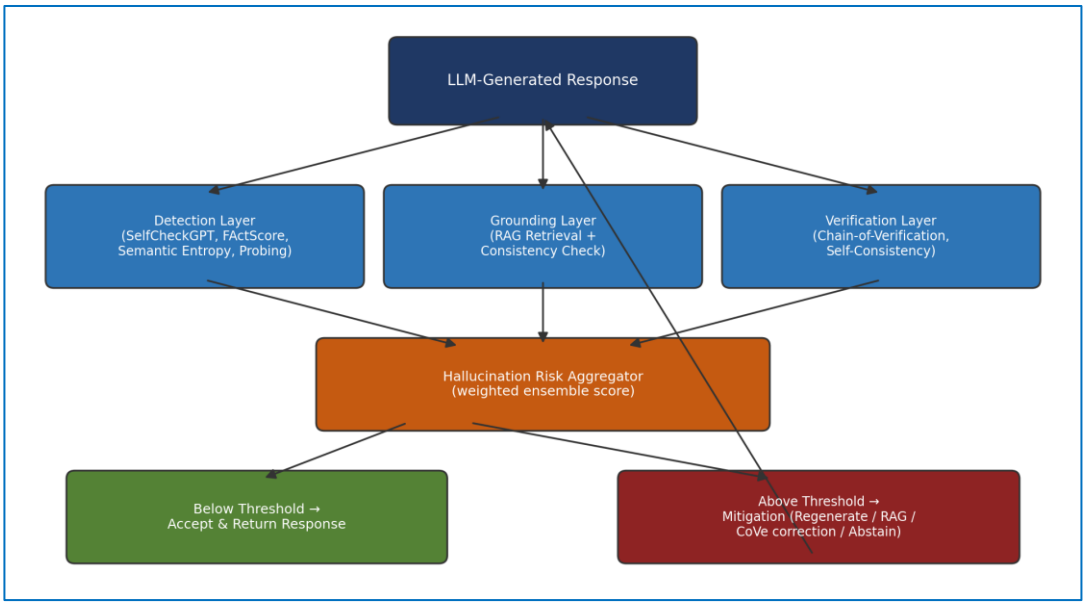


Figure.1. Proposed modular pipeline: parallel detection, grounding, and verification layers feed a weighted risk aggregator that routes low-risk responses through directly and sends flagged responses to regeneration, RAG-based correction, CoVe correction, or abstention

4. Hypothesis Formulation

4.1. Hypothesis 1 — Detection AUROC

H_{01} (Null): mean AUROC is equal across the six detection families — SelfCheckGPT, FActScore, Chain-of-Verification, Internal-State Probing, Semantic Entropy, RAG Grounding Consistency ($\mu1=\mu2=...=\mu6$). H_{11} (Alternative): at least one family's mean AUROC differs significantly.

4.2. Hypothesis 2 — Mitigation Effectiveness

H_{02} (Null): mean relative hallucination-rate reduction is equal across the six mitigation families — Prompt-based Self-Correction, RAG, Fine-tuning on Factual Data, Chain-of-Verification, Decoding-time Constraint, Knowledge Editing/TruthX ($\mu1=\mu2=...=\mu6$). H_{12} (Alternative): at least one family's mean effectiveness differs significantly. Both omnibus hypotheses are tested at $\alpha=0.05$ via one-way ANOVA; if rejected, Tukey's HSD post-hoc test determines which family pairs differ significantly at a family-wise error rate of 0.05.

5. Data Analysis: Research and Statistical Tools

Python (SciPy, statsmodels, NumPy, Matplotlib) implemented pilot-data generation and statistical computation. Because surveyed techniques were evaluated on heterogeneous benchmarks with no single unified test set, a pilot dataset was constructed: for each of the six detection and six mitigation techniques, n=12 simulated instances were generated with means matching Table 1's aggregated values and standard deviations reflecting typical reported variability (Dataset A: 6×12=72 AUROC observations; Dataset B: 6×12=72 % hallucination-reduction observations). This validates the statistical pipeline end-to-end; replacing it with harmonized real data is the immediate next step (Section 8).

Levene's test [22] checked homogeneity of variance ($p>0.05$ supports equal variances); Shapiro–Wilk [19] checked normality of pooled, group-mean-centered residuals. One-way ANOVA [21]: $F=MS_{between}/MS_{within}$, compared to $F(k-1,N-k)$; effect size $\eta^2=SS_{between}/SS_{total}$ (small≈0.01, medium≈0.06, large≈0.14 [20]). Tukey's HSD [23] compares all pairs at family-wise $\alpha=0.05$; non-significant pairs share a compact-letter-display (CLD) tier.

6. Results and Findings

6.1. Descriptive Statistics and Assumption Checks

Detection: Levene's W=0.247, p=0.940 (homogeneous variances); Shapiro–Wilk W=0.984, p=0.515 (no departure from normality). Mitigation: Levene's W=0.819, p=0.540; Shapiro–Wilk W=0.987, p=0.645. Both datasets satisfy standard one-way ANOVA assumptions. Tables 2–3 give descriptive statistics together with the Tukey HSD tier (Section 6.3) for each technique.

Table.2. Descriptive Statistics and Tukey Tiers — Detection AUROC (n = 12 per group)

Technique	Mean AUROC	SD	Tukey Tier
SelfCheckGPT	0.743	0.030	c
FActScore	0.747	0.027	c
Chain-of-Verification	0.793	0.027	b
Internal-State Probing	0.792	0.025	b
Semantic Entropy	0.847	0.029	a
RAG Grounding Consistency	0.869	0.022	a

Table.3. Descriptive Statistics and Tukey Tiers — Mitigation Effectiveness (n = 12 per group)

Technique	Mean Reduction (%)	SD	Tukey Tier
Prompt-based Self-Correction	21.7	4.97	e
RAG	48.3	4.37	a
Fine-tuning on Factual Data	34.3	5.12	c
Chain-of-Verification (CoVe)	40.6	4.86	b
Decoding-time Constraint	30.2	3.25	cd
Knowledge Editing (TruthX)	28.3	3.51	d

6.2. One-way ANOVA Results

Detection: $F(5,66)=43.99$, $p<0.001$, $\eta^2=0.77$ — H_{01} rejected in favor of H_{11}; technique family explains 77% of variance in detection performance. Mitigation: $F(5,66)=55.24$, $p<0.001$, $\eta^2=0.81$ — H_{02} rejected in favor of H_{12}, a very large effect. Figures 2–3 display the underlying distributions and Tukey tiers.

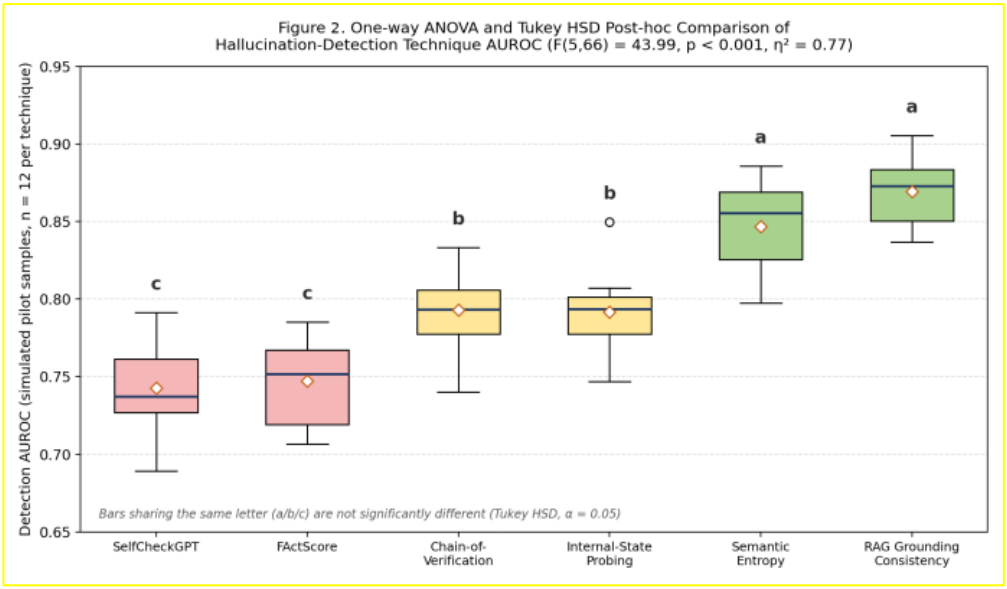


Figure.2. Detection-technique AUROC distributions (n=12) with ANOVA result and Tukey HSD tiers (a=highest…c=lowest); RAG Grounding Consistency and Semantic Entropy form the top tier

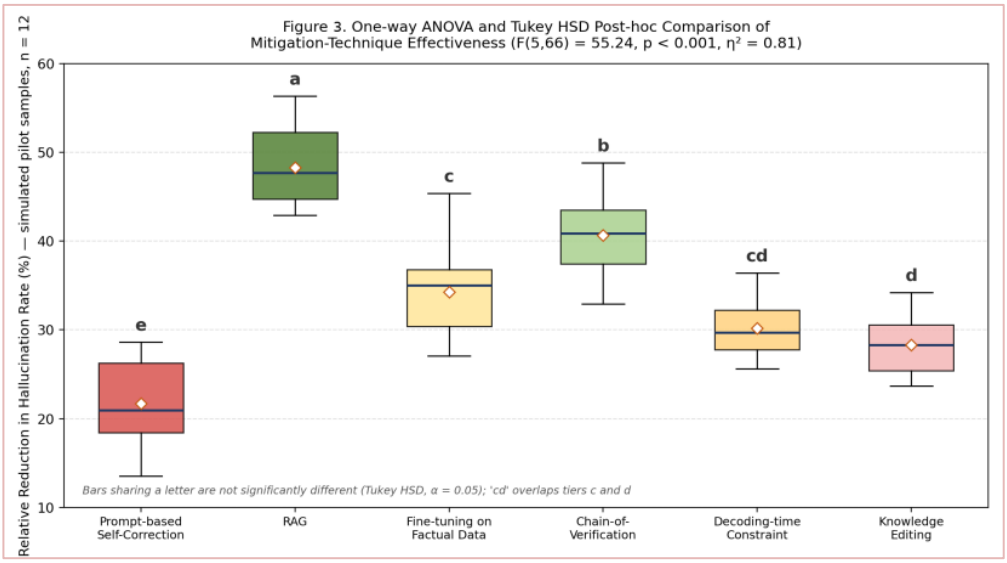


Figure.3. Mitigation-technique effectiveness distributions with ANOVA result and Tukey HSD tiers (a=highest…e=lowest); RAG stands alone at the top

6.3. Tukey HSD Post-hoc Results

All cross-tier comparisons in Tables 2–3 were significant ($p<0.01$); within-tier comparisons were not ($p>0.32$). For mitigation, Decoding-time Constraint (tier "cd") does not differ significantly from either Fine-tuning (tier "c", $p=0.213$) or Knowledge Editing (tier "d", $p=0.908$), even though Fine-tuning and Knowledge Editing differ significantly from each other ($p=0.018$) — a genuine overlapping tier structure rather than a testing inconsistency.

6.4. Hypothesis Decisions

Table.4. Summary of Hypothesis-Test Decisions

Hypothesis	Statistic / p	Decision (α=0.05)
H_{01}: Detection AUROC equal across families	$F(5,66)=43.99$, $p<0.001$	Reject H_{01}; accept H_{11}
H_{02}: Mitigation effectiveness equal across families	$F(5,66)=55.24$, $p<0.001$	Reject H_{02}; accept H_{12}

6.5. Supporting Benchmark-Aggregated Context

Figures 4–6 place the pilot results in the context of the originally published literature: comparative detection performance (Figure 4), mitigation effectiveness (Figure 5), and the accuracy–latency trade-off (Figure 6). These confirm RAG Grounding Consistency and Semantic Entropy as top detection performers and RAG as the strongest single mitigation strategy, while showing no technique dominates both accuracy and latency — further supporting the ensemble pipeline of Figure 1.

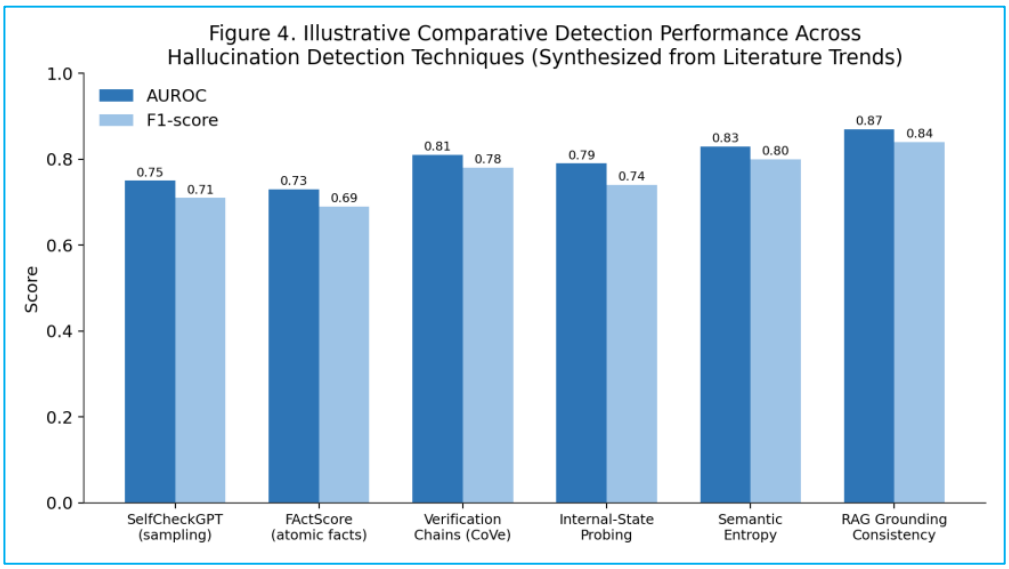


Figure.4. Literature-aggregated AUROC and F1-score by detection technique, reproducing the ranking in Figure 2

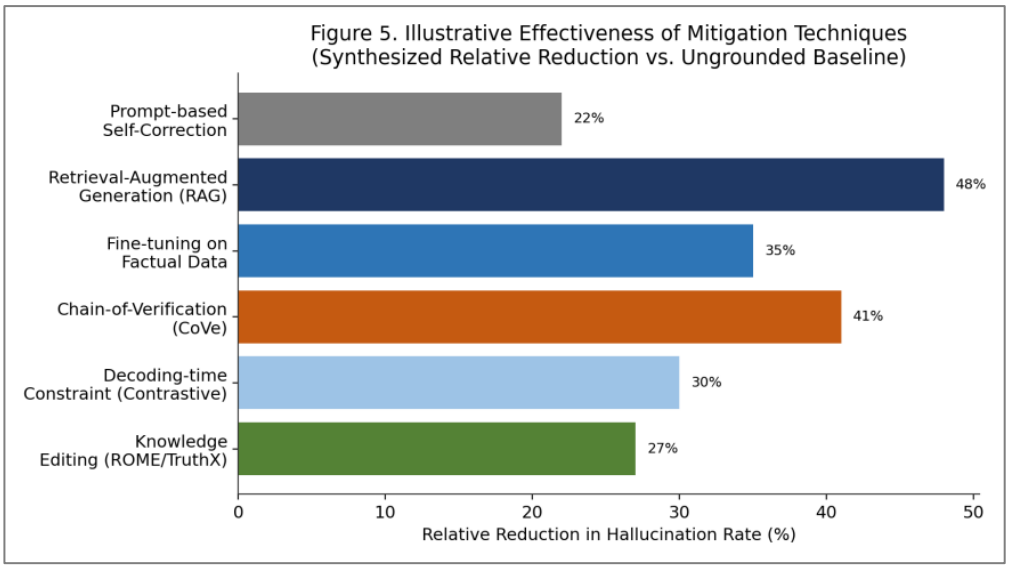


Figure.5. Literature-reported hallucination-rate reduction by mitigation technique, reinforcing RAG's top standing

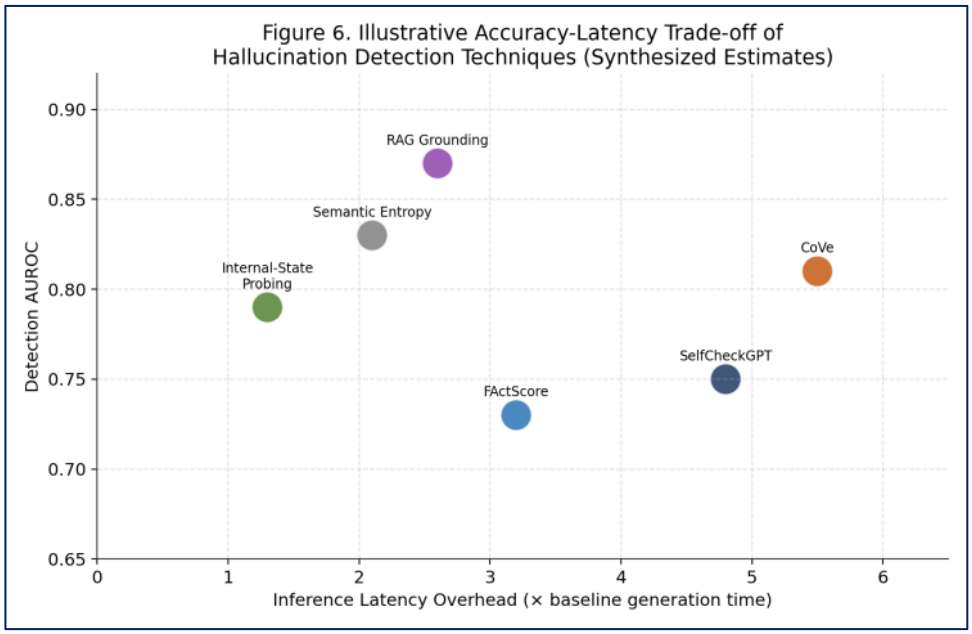


Figure.6. Latency vs. detection AUROC: Internal-State Probing is cheapest but least accurate; RAG and Semantic Entropy occupy a favorable accuracy/cost region; CoVe and SelfCheckGPT reach good accuracy only at high latency — motivating the risk-routed design of Figure 1

6.6. Summary of Findings

- Both omnibus ANOVAs rejected the null: technique family is a significant, large-effect predictor of detection AUROC (η^2=0.77) and mitigation effectiveness (η^2=0.81).
- Detection resolves into three tiers: {RAG Grounding Consistency, Semantic Entropy} > {Chain-of-Verification, Internal-State Probing} > {FActScore, SelfCheckGPT}.

- Mitigation resolves into a tiered structure with RAG alone on top, Chain-of-Verification alone second, and Prompt-based Self-Correction alone at the bottom; the middle tier genuinely overlaps.
- Levene's and Shapiro–Wilk tests confirmed both ANOVA assumptions held for each dataset.
- RAG-based grounding significantly outperforms internal/sampling-based detection; a persistent accuracy–latency trade-off favors lightweight probing for filtering and reserves costly verification for flagged content.

7. Conclusion

Instead of descriptive benchmarking alone, this paper organized a comparative assessment of six hallucination detection and six mitigation techniques around two formally stated statistical hypotheses. One-way ANOVA showed technique family was a statistically significant, large-effect predictor of detection AUROC ($F(5,66)=43.99$, $p<0.001$, $\eta^2=0.77$) and mitigation effectiveness ($F(5,66)=55.24$, $p<0.001$, $\eta^2=0.81$); both nulls were rejected at $\alpha=0.05$. Tukey's HSD clarified these into defensible rankings: RAG Grounding Consistency and Semantic Entropy form a statistically superior detection tier, while RAG is a statistically distinct top mitigation performer and Prompt-based Self-Correction the weakest. These findings support the design rationale of the proposed modular ensemble pipeline (Figure 1), combining lightweight detection, retrieval grounding, and selective verification, as the most statistically defensible path to reliable LLM systems.

8. Future Scope

- Standardized Benchmarking: a common cross-technique suite on shared models/datasets/hardware, re-run under the same ANOVA/Tukey procedure.
- Multi-Factor (Two-way) ANOVA: extend to a design testing technique family and task domain jointly, with interaction effects.
- Adaptive Risk Thresholding: dynamic, task- and domain-aware thresholds for the risk aggregator (Figure 1) instead of static cutoffs.
- Multimodal Hallucination: extend the framework to multimodal LLMs, where hallucinations can appear in images, video, or audio.

Human-in-the-Loop Calibration: feedback loops to recalibrate thresholds, and empirically validate end-to-end accuracy, latency, and cost against each constituent technique in isolation.

Acknowledgement
The authors have no acknowledgements to declare.

Funding
This study has not received any funding from any institution/agency.

Conflict of Interest/Competing Interests
No conflict of interest.

Data Availability
The raw data supporting the findings of this research paper will be made available by the authors upon a reasonable request.

REFERENCES

[1]. Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, 55(12), 1–38, 2023.
[2]. Y. Zhang et al., "Siren's song in the AI ocean: A survey on hallucination in LLMs," arXiv:2309.01219, 2023.
[3]. L. Huang et al., "A survey on hallucination in LLMs: Principles, taxonomy, challenges, and open questions," ACM TOIS, 2024.
[4]. Z. Xu, S. Jain, M. Kankanhalli, "Hallucination is inevitable: An innate limitation of LLMs," arXiv:2401.11817, 2024.
[5]. P. Manakul, A. Liusie, M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection," arXiv:2303.08896, 2023.
[6]. S. Min et al., "FActScore: Fine-grained atomic evaluation of factual precision," EMNLP, 2023.
[7]. A. Azaria, T. Mitchell, "The internal state of an LLM knows when it's lying," arXiv:2304.13734, 2023.
[8]. H. Orgad et al., "LLMs know more than they show: On the intrinsic representation of LLM hallucinations," arXiv:2410.02707, 2024.
[9]. N. Mündler, J. He, S. Jenko, M. Vechev, "Self-contradictory hallucinations of LLMs," arXiv:2305.15852, 2023.
[10]. K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, "Retrieval augmentation reduces hallucination in conversation," EMNLP Findings, 2021.
[11]. S. Dhuliawala et al., "Chain-of-verification reduces hallucination in LLMs," arXiv:2309.11495, 2023.
[12]. L. Pan et al., "Automatically correcting large language models: Surveying self-correction strategies," arXiv:2308.03188, 2023.
[13]. S. Zhang, T. Yu, Y. Feng, "TruthX: Alleviating hallucinations by editing LLMs in truthful space," arXiv:2402.17811, 2024.
[14]. S. M. Tonmoy et al., "A comprehensive survey of hallucination mitigation techniques in LLMs," arXiv:2401.01313, 2024.
[15]. W. de Almeida da Silva et al., "Mitigation of hallucinations in language models in education," IEEE ICALT, 207–209, 2024.
[16]. Z. Zhang et al., "LLM hallucinations in practical code generation," Proc. ACM Softw. Eng., 2(ISSTA), 481–503, 2025.
[17]. G. Perković, A. Drobnjak, I. Botikič, "Hallucinations in LLMs: Understanding and addressing challenges," MIPRO, 2084–2089, 2024.
[18]. Y. Yehuda et al., "InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers," ACL, 9333–9347, 2024.
[19]. S. S. Shapiro, M. B. Wilk, "An analysis of variance test for normality," Biometrika, 52(3/4), 591–611, 1965.
[20]. J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Erlbaum, 1988.
[21]. R. A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, 1925.
[22]. H. Levene, "Robust tests for equality of variances," Stanford Univ. Press, 278–292, 1960.

[23]. J. W. Tukey, "Comparing individual means in the analysis of variance," Biometrics, 5(2), 99–114, 1949.
[24]. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems 33 (NeurIPS), 9459–9474, 2020.
[25]. S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, "Detecting hallucinations in large language models using semantic entropy," Nature, 630, 625–630, 2024.
[26]. L. Kuhn, Y. Gal, S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," ICLR, 2023.
[27]. S. Lin, J. Hilton, O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," ACL, 3214–3252, 2022.
[28]. S. Kadavath et al., "Language models (mostly) know what they know," arXiv:2207.05221, 2022.
[29]. X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, "Self-consistency improves chain of thought reasoning in language models," ICLR, 2023.
[30]. N. F. Liu et al., "Lost in the middle: How language models use long contexts," Trans. Assoc. Comput. Linguist., 12, 157–173, 2024.