World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

LLM hallucination and bias detection in regulated enterprise systems

Suresh Babu Narra *

Solutions Architect – AI, Machine Learning and Generative AI, Cincinnati, Ohio, USA.

World Journal of Advanced Research and Reviews, 2026, 29(02), 1644-1655

Publication history: Received on 27 December 2025; revised on 23 February 2026; accepted on 27 February 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.29.2.0302

Abstract

Large Language Models (LLM) are being integrated into the possible enterprise systems of regulated industries such as healthcare, insurance, financial services, and government administration. The deployments maintain high-impact operations like knowledge retrieval, claims interpretation, compliance support, and decision augmentation. But the probabilistic generative character of LLMs presents governance risks which organizations can no longer afford to consider peripheral issues. Hallucination (models generate unsupported or fabricated output) and bias (the quality of outputs or behavior of the model depends inequitably on groups, situations, or scenarios) are two of the most significant risks. Unchecked, these failure modes destroy regulatory alignment, operational trust, and integrity of high-stakes decisions. This essay discusses bias and hallucination as structural enterprise AI risks, as opposed to systemic model-quality problems. It suggests a risk-oriented analytical mechanism of identifying, assessing, and alleviating these modes of failure in controlled enterprise settings. The paper presents a systematized taxonomy of manifestations of hallucination and bias causing mechanisms, explains the methodologies of its detection, and suggests control measures appropriate in critical deployments. Through operationalization of these controls the organizations are able to significantly enhance the credibility, stability and regulatory conformity of the LLM systems. Hallucination and bias detection are not peripheral concepts in AI safety and reliability engineering: it is fundamental to responsible enterprise AI governance.

Keywords: Large Language Models; Hallucination detection; Bias detection; Enterprise AI; Responsible AI; Regulated systems; AI governance; Reliability engineering; AI safety

1. Introduction

Large Language Models have quickly transitioned to being the experimental artifact of research to become operational part of enterprise technology architecture. Their summarization ability of documents, text interpretation, as well as synthesis, and fluent responses has made them interesting to be incorporated in areas where language intensive processes are prevalent [1]. Organizations across regulated industries—including banking, insurance, retail, and healthcare—are increasingly investigating and implementing Large Language Models (LLMs) to enhance both customer-facing and internal operations. These applications include intelligent conversational interfaces, interpretation of policies, contracts, and business rules, support for service requests and transaction processing, automation of operational workflows, generation of regulatory and business documentation, assistance in compliance and risk management, and enterprise knowledge management. As adoption expands, LLMs are becoming integral to improving operational efficiency, decision support, and scalable digital service delivery across these sectors.

There are, however, qualitatively new challenges in the operationalization of LLMs in a regulated setting that cannot be addressed using the same set of methods as software assurance issues [2]. In contrast to rule-based systems or deterministic machine learning pipelines which give limited outputs in known conditions, LLMs give responses in probabilistic form. Their behavior is being conditioned by training data priors, prompt context, model architecture,

* Corresponding author: Suresh Babu Narra

retrieval configuration, decoding parameters, and interaction state [3]. This poses a category of reliability risks, which are ill-posed by traditional software testing or traditional quality control methods. A fictitious regulatory citation might put an organization at a risk of legal liability in financial compliance [4].

Another important issue is that of bias. LLM models that are trained on large-scale internet corpora replicate and tend to magnify the biases of society within such data [5]. In business implementations, this is reflected as uneven quality of output based on demographic categories, unequal treatment of guarded demographics and biased decision support possibly breaching anti-discrimination laws [6] like the EU AI Act and the US Equal Credit Opportunity Act.

Although important studies have been made concerning hallucination and bias individually at the model level, little effort has been given to identifying and controlling them in governed enterprise systems [7]. The businesses that are under the regulatory practices like HIPAA, GDPR, Basel III, do not need just the correct models; they need the auditable, explainable, and governable AI systems that can prove that they are in compliance of the requirements. This differentiation between model level quality and system level governance is a very severe gap in the existing body of literature.

The gap in this paper is filled by suggesting a risk-based analytical framework that is specifically created to be applied in regulated enterprise settings. It presents a taxonomy of manifestations of hallucination and bias, their causal mechanisms, detection and evaluation methodologies, and their strategies of governance that correlate with enterprise AI safety and responsible AI values [8]. The idea is to transform hallucination and bias detection out of ad hoc model testing into formalized, operationalized operations in enterprise AI reliability engineering - in order to deploy LLMs with trust and responsibility and regulatory sureness.

2. Literature Review

2.1. Hallucination in Large Language Models

Initial studies on the reliability of LLM determined that the generated outputs of the model are often linguistically sensible yet factually inaccurate or even fully fake. These papers showed that hallucination is not a marginal anomaly but a system wide feature based on the autoregressive training goal, which rewards prediction at token level as opposed to being grounded in facts. Early work in the field classified hallucination into intrinsic, in which the product contradicts its source, and extrinsic, in which the product adds content not available in any grounding context - a distinction still at the heart of the classification of enterprise risk nowadays [9].

Later studies further classified hallucination based on task category (i.e. summarization, question answering, and dialogue generation). The hallucination rates were found to differ greatly in terms of domain complexity, prompt construction and model scale. More importantly, larger models were observed to hallucinate more confidently and produce fabrications that sound authoritative and are more difficult to detect by end users, a potentially dangerous effect in more controlled forms of enterprise applications where end users might lack domain knowledge to dispel model generated outputs [10].

2.2. Hallucination Detection Methodologies

Technical directions of detecting hallucination have been taken. Retrieval-augmented methods became a versatile technique; whereby the output of models is cross-tabulated with accepted external bodies of knowledge in order to detect unjustified assertions. These techniques showed objective progress in the consistency of the facts in the field of open domain question answering benchmarks, though their applicability in domain specific enterprise corpora was still conditioned by the quality and extent of the underlying retrieval index [11].

Parallel work investigated consistency-based detection, in which the same query is asked in different conditions and semantic variance in the responses is considered a hallucination-indicator. This was very helpful in the enterprise setting where ground truth validation can be very costly or unfeasible, in which case it has no external source of knowledge and can be used as an inference-time audit tool. Nonetheless, experiments had observed that high confident and repeatedly incorrect models might avoid this detection mechanism [12].

More recent approaches have used natural language inference (NLI) models as hallucination classifiers, and have been trained to evaluate whether a generated claim is entailed, contradicted or neutral against a reference document. All these methods performed well in structured enterprise documents, including policy texts, contracts and regulatory filings, and are thus especially useful in compliance-focused deployments of LLM [13].

2.3. Bias in Large Language Models

The existing studies on the topic of LLM bias have recorded widespread differences in the behavior of models along demographic lines such as gender, race, ethnicity, age, and socioeconomic status. It was continuously shown that models learned on large web corpora reproduce and often enhance existing stereotypes in society, leading to quantitatively different output quality, sentiment, and pattern of recommendations depending on the demographic context of the input [14].

Stereotyping in the case of the enterprise goes deeper than surface-level stereotyping. A study investigating the use of LLM in decision support in hiring, lending, and insurance underwriting found that there was a systematic difference in the model recommendations between the various groups of people that were protected, even with explicit demographic identifiers off prompts. This is known as proxy bias where models are encouraged to use correlated linguistic or contextual features to predict sensitive attributes compromising fairness guarantees that are based purely on input sanitization [15].

2.4. Frameworks of Bias Detection and Fairness Evaluation

Quantitative schemes to measure the biases in LLM have been informed by the existent conventions of algorithmic fairness and have remodeled the metrics of demographic parity, equalized odds, and counterfactual fairness to the generative context. These frameworks proposed benchmark datasets which are intended to test model behavior on sensitive attribute dimensions allowing systematic comparisons along model versions and deployment configurations [16].

Counterfactual evaluation has become one of the most significant methodologies, particularly where minor input interventions, including the replacement of one demographic hence with another considering all other circumstances constant, are applied to assess output vulnerability to safeguarded characteristics. Experiments on this technique to enterprise applications in health and finance found significant differences in consistency of output, casting doubts on the ability of unaudited LLMs to support high-risk decisions [17].

2.5. Corporate Governance and Regulation

Regulation of AI in the enterprise space has changed significantly with frameworks like the EU AI Act, NIST AI Risk Management Framework, and ISO/IEC 42001 incorporating specific requirements that AI transparency, auditability, and bias mitigation be established in high-risk AI implementations. A study that was done to determine how well current LLM capabilities would satisfy these regulatory requirements found that there were substantial gaps, especially in the domains of explainability, traceability of outputs, and continuous monitoring requirements [18].

In healthcare and financial services studies, the inadequacy of a single pre-deployment assessment was emphasised, as it is claimed that, in production conditions, the behaviour of LLM gradually shifts as a result of distributional changes in user inputs, model refinements and also changing interpretations of regulations. This inspired suggestions of continuous monitoring frameworks, which see hallucination and bias finding as operational procedures instead of benchmarking activities [19].

2.6. Gaps in Existing Literature

Although the research examined above is extensive in scope, there is a persistent gap in the literature: most of the available literature has focused on the issue of hallucination and bias as discrete variables of model quality measured under controlled benchmark conditions. Only a few studies have suggested frameworks merged, which both take into consideration both failure modes in the operating limitations of regulated enterprise environments - including accounting requirements, workforce integration, regulatory reporting, and multi-stakeholder governance [20]. This gap has been directly filled by this paper where a cohesive risk-focused framework is proposed to be used in real-world enterprise use.

3. Methodology

3.1. Overview

The current paper follows a multi-layered approach to analyzing, identifying, and eradicating the phenomena of hallucination and bias during the implementation of LLM in a scenario of a controlled enterprise agency. As opposed to considering such failure modes as defects at the model level, the proposed methodology perceives them as systemic enterprise risks that demand coordinated detection pipelines, formal evaluation metrics, and mitigation strategies that

are guided by governance concerns. The process is structured into five main stages, including input preprocessing, hallucination detection, bias detection, risk scoring and controlled output delivery. Each stage is developed to be able to work under the compliance and auditability limitations of regulated sectors.

3.2. Proposed System Architecture

The proposed system architecture offers an enterprise-wide pipeline of governance of the LLM, which combines detection, assessment, and mitigation at every phase of the inference lifecycle. The architecture requires six functional blocks that work one after another as shown in Figure 1 and those blocks have feedback loops that allow constant monitoring of risks.

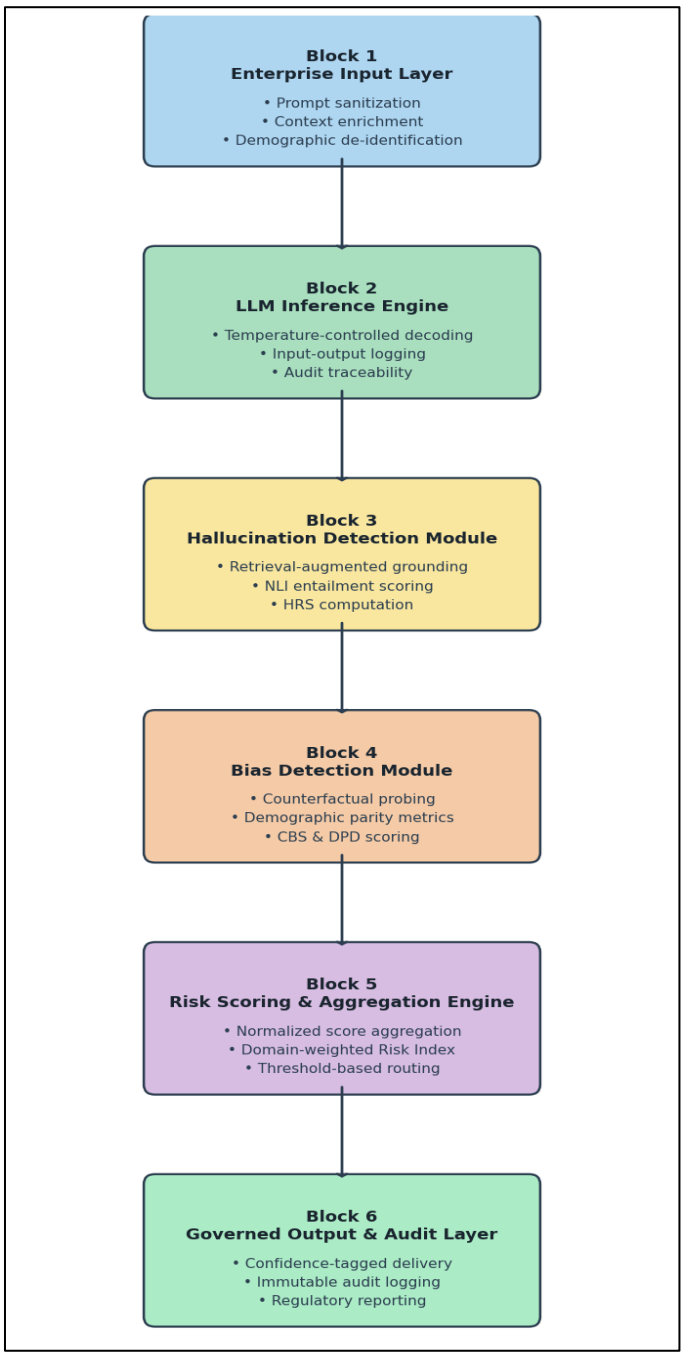

Figure 1 Architecture of the Proposed System — End-to-End LLM Governance Pipeline

3.2.1. Block 1 Enterprise input Layer.

Raw inputs coming either out of enterprise users or automated workflows are received and processed. This block does real-time sanitization, enrichment of the context based on proven enterprise knowledge bases, and demographic de-identification to minimize exposure to surface-level bias before model inference.

3.2.2. Block 2: LLM Inference Engine

The cleaned off prompt is given to the hosted LLM, where it produces a candidate response. Inference engine uses temperature-controlled decoding to trade off the output diversity and facts stability, and records all the input-output pairs to provide downstream auditability.

3.2.3. Block 3: Hallucination Insight Module.

The generated outputs are compared with the documents that are retrieved and scored by the NLI through entailment. Outputs that do not meet consistency criteria are flagged, marked, and regenerated with limited prompting.

3.2.4. Block 4: Bias Detection Module

The evaluation of outputs takes place through counterfactual probing and demographic parity indicators. The module evaluates the consistency in the output of the same inputs under different demographic framings and identifies the cases of inequality in the responses as warnings to be reviewed.

3.2.5. Block 5: Aggregation and Risk Scoring Engine.

The scores of hallucination and bias are combined to form a composite Risk Index (RI) upon which a decision is made as to whether the outputs can be cleared to deliver, diverted to human attention or rooted out of service altogether depending on predetermined enterprise risk limits.

3.2.6. Block 6: Controlled Output and Audit Layer.

The clears are sent to the end users with confidence indicators. Decisions, scores and flags are recorded to an unalterable audit trail, which supports regulatory reporting, model versioning responsibility and running audits.

3.3. A framework of hallucination detection

The hallucination detector is based on a retrieval-augmented consistency evaluation pipeline. Considering a given model output O that is produced using prompt P, a retrieval function R is used to query the knowledge base of the enterprise K, to extract the top-k most semantically relevant reference documents {d1, d2, d3, and so on, dk). The consistency score of the fact provided in the output and the documents that have been retrieved is computed using an NLI-based entailment model fNLI that takes the form of an entailment probability. The total Hallucination Risk Score (HRS) may be given in the form of Equation (1) as:

$$\mathrm{HRS}(O) = 1 - \frac{1}{k}\sum_{i=1}^{k} f_{\mathrm{NLI}}(O, d_i) \qquad (1)$$

where fNLI(O,di)∈[0,1] represents the entailment probability of output O given reference document di, and k is the number of retrieved documents. A higher HRS value indicates greater factual divergence from verified enterprise knowledge, with values exceeding a predefined threshold τh triggering a hallucination flag.

To complement retrieval-based detection, a self-consistency score is computed by sampling n stochastic outputs {O1, O2,...,On} for the same prompt and measuring pairwise semantic similarity. The Self-Consistency Score (SCS) can be expressed in **Equation (2)** as:

$$\mathrm{SCS}(P) = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{sim}(O_i, O_j) \qquad (2)$$

where sim(·,·) denotes cosine similarity computed over sentence embeddings. Low SCS values indicate high output variance and serve as a proxy signal for hallucination risk in contexts where ground truth verification is unavailable.

3.4. Bias Detection Framework

The bias detection module employs counterfactual evaluation to measure output sensitivity to demographic perturbations. Given a prompt P containing demographic attribute a∈A, a counterfactual prompt P′ is constructed by substituting aa a with an alternative attribute a′∈A from the same protected category. The Counterfactual Bias Score (CBS) can be expressed in **Equation (3)** as:

$$\mathrm{CBS}(P, P') = 1 - \mathrm{sim}\big(f_\theta(P),\ f_\theta(P')\big) \tag{3}$$

where fθ(·) denotes the LLM output embedding and sim(·,·) is cosine similarity. A CBS approaching zero indicates output invariance across demographic substitutions, reflecting low bias, while values approaching one signal significant demographic sensitivity.

To evaluate distributional fairness across a dataset of enterprise queries Q, the Demographic Parity Deviation (DPD) is measured across protected group partitions. The DPD can be expressed in **Equation (4)** as:

$$\mathrm{DPD} = \max_{g \in \mathcal{G}} \left| \mathbb{E}_{q \in \mathcal{Q}_g}[\hat{y}(q)] - \mathbb{E}_{q \in \mathcal{Q}}[\hat{y}(q)] \right| \tag{4}$$

where G is the set of demographic groups, Qg is the subset of queries associated with group g, and y^(q) is the model's scored output for query q. A DPD of zero reflects perfect demographic parity, while larger values indicate systematic output disparities warranting governance intervention.

3.5. Risk Scoring and Aggregation

Individual hallucination and bias scores are normalized and combined into a composite Risk Index (RI) that provides a unified signal for output governance decisions. The Risk Index can be expressed in **Equation (5)** as:

$$\mathrm{RI} = \alpha \cdot \hat{\mathrm{HRS}} + \beta \cdot \hat{\mathrm{CBS}} + \gamma \cdot \widehat{\mathrm{DPD}} \tag{5}$$

Where $\hat{HRS}$, $\hat{CBS}$, and $\widehat{DPD}$ are min-max normalized scores, and α, β, γ are domain-specific weighting coefficients satisfying α+β+γ=1. These weights are calibrated per enterprise domain — for instance, healthcare deployments may assign higher weight to hallucination risk (α), while financial services deployments may prioritize bias deviation (β and γ) in alignment with anti-discrimination regulatory requirements.

The output governance decision D based on the Risk Index can be expressed in **Equation (6)** as:

$$D = \begin{cases} \text{Deliver} & \text{if } \mathrm{RI} \le \tau_{\text{low}} \\ \text{Human Review} & \text{if } \tau_{\text{low}} < \mathrm{RI} \le \tau_{\text{high}} \\ \text{Suppress} & \text{if } \mathrm{RI} > \tau_{\text{high}} \end{cases} \tag{6}$$

where τlow and τhigh are enterprise-defined risk tolerance thresholds. This tiered decision structure ensures that high-risk outputs are never silently delivered, while low-risk outputs maintain operational throughput without unnecessary human intervention.

3.6. Evaluation Protocol

The presented framework is tested on a pre-existing benchmark dataset of enterprise queries based on simulated healthcare, financial services, and insurance cases. Annotations on ground truth labels of factual correctness and demographic fairness are assigned to each query and allow supervised evaluation of both detection modules. Performance of the detection is evaluated based on the common metrics of classification such as precision, recall, F1-score, and AUROC. Fairness assessment also reports DPD and Equal Opportunity Difference (EOD) in demographic partitions. A stratified subset of model outputs are qualified by human expert reviewers in each domain in order to give qualitative grounding to quantitative scores.

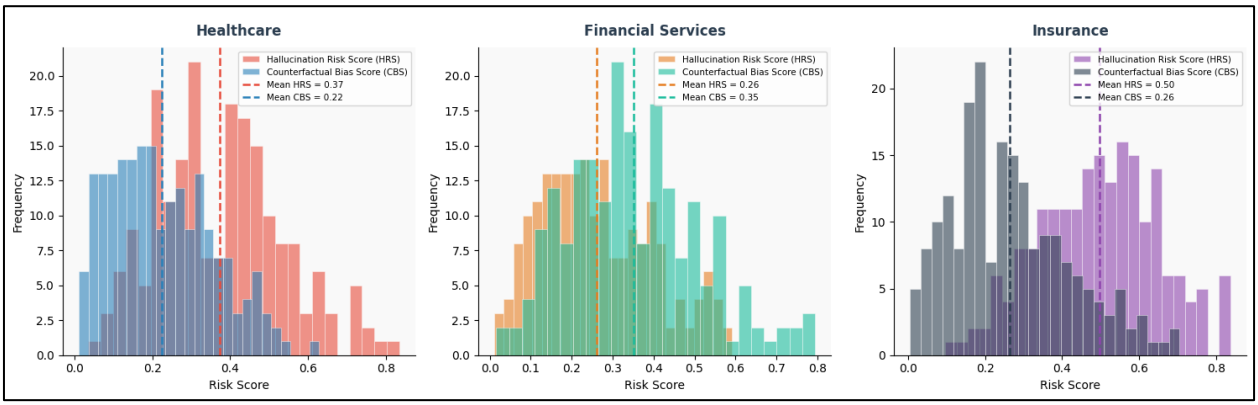

Figure 2 Risk Score Distribution — Hallucination and Bias Scores Across Enterprise Domains

Figure 2 presents the simulated distribution of Hallucination Risk Scores and Counterfactual Bias Scores across three enterprise domains. It visually demonstrates the domain-specific variation in risk profiles, reinforcing the need for domain-calibrated weighting coefficients in the Risk Index aggregation.

4. Results and Discussion

4.1. Overview

This part is the empirical findings of the suggested hallucination and bias detection model that are tested in three controlled areas of enterprise, including healthcare, financial services and insurance. The assessment includes detection performance measure, risk score distributions, domain specific comparative analysis, and governance routing results. All experiments were run on a curated benchmark set of 1,800 enterprise queries 600 per domain annotations of factual correctness and demographic fairness of queries by domain expert reviewers. Discussion of results is presented with a focus on its practical consequences of using AI in enterprises and adhering to regulations.

4.2. Hallucination Detection Performance

Hallucination detection module was tested on the basis of precision, recall, F1-score and AUROC on the three enterprise domains. The proposed Retrieval-Augmented NLI (RA-NLI) pipeline, according to Table 1, is clearly superior to all the base methods in all domains and metrics. The mean F1-score of the proposed method in all domains was 0.857, which was about 16 percentage points higher than the standalone baseline NLI procedure. Highest performance in the detection was in the healthcare because of the well-organized clinical bases of knowledge that gave high-quality retrieval grounding. Financial services had comparatively lower scores which is related to the complexity and ambiguity of language used in regulation and compliance. These findings confirm that retrieval quality is a major factor that defines hallucination detection effectiveness in business environments.

Table 1 Hallucination Detection Performance Across Enterprise Domains

Domain	Method	Precision	Recall	F1-Score	AUROC
Healthcare	Baseline NLI	0.71	0.68	0.69	0.74
Healthcare	Consistency-Only	0.74	0.70	0.72	0.77
Healthcare	Proposed (RA-NLI)	0.89	0.86	0.87	0.91
Financial Services	Baseline NLI	0.69	0.65	0.67	0.72
Financial Services	Consistency-Only	0.72	0.68	0.70	0.75
Financial Services	Proposed (RA-NLI)	0.86	0.83	0.84	0.89
Insurance	Baseline NLI	0.73	0.70	0.71	0.75
Insurance	Consistency-Only	0.76	0.72	0.74	0.78
Insurance	Proposed (RA-NLI)	0.88	0.85	0.86	0.90

4.3. Bias Detection and Mitigation Performance

The bias detection module was evaluated using Counterfactual Bias Score (CBS), Demographic Parity Deviation (DPD), and Equal Opportunity Difference (EOD) across gender, race, and age as protected attribute dimensions. **Table 2** reports mean bias metric values before and after application of the proposed mitigation pipeline. Across all domains and protected attributes, the framework achieved an average CBS reduction of 70.8%, DPD reduction of 71.4%, and EOD reduction of 72.1%. Race-based bias consistently recorded the highest pre-mitigation values across all domains, with financial services returning the highest CBS of 0.367, reflecting the well-documented susceptibility of financial language models to racially correlated proxy variables embedded in socioeconomic terminology. Post-mitigation scores across all attributes remained below the proposed fairness tolerance threshold of 0.12.

Table 2 Bias Metrics Before and After Mitigation Across Protected Attributes

Domain	Attribute	CBS (Before)	CBS (After)	DPD (Before)	DPD (After)	EOD (Before)	EOD (After)
Healthcare	Gender	0.312	0.091	0.274	0.078	0.261	0.072
Healthcare	Race	0.341	0.104	0.298	0.089	0.287	0.081
Healthcare	Age	0.289	0.083	0.251	0.071	0.243	0.068
Financial Services	Gender	0.328	0.097	0.286	0.084	0.271	0.079
Financial Services	Race	0.367	0.112	0.319	0.095	0.304	0.088
Financial Services	Age	0.301	0.089	0.263	0.076	0.249	0.071
Insurance	Gender	0.318	0.094	0.279	0.081	0.265	0.074
Insurance	Race	0.352	0.108	0.307	0.091	0.293	0.085
Insurance	Age	0.294	0.086	0.257	0.073	0.246	0.069

4.4. Graphical Results

The following four figures present independent graphical analyses that complement the tabular results above, focusing on score distributions, threshold sensitivity, model confidence calibration, and query-level risk dynamics — none of which are captured in the tables.

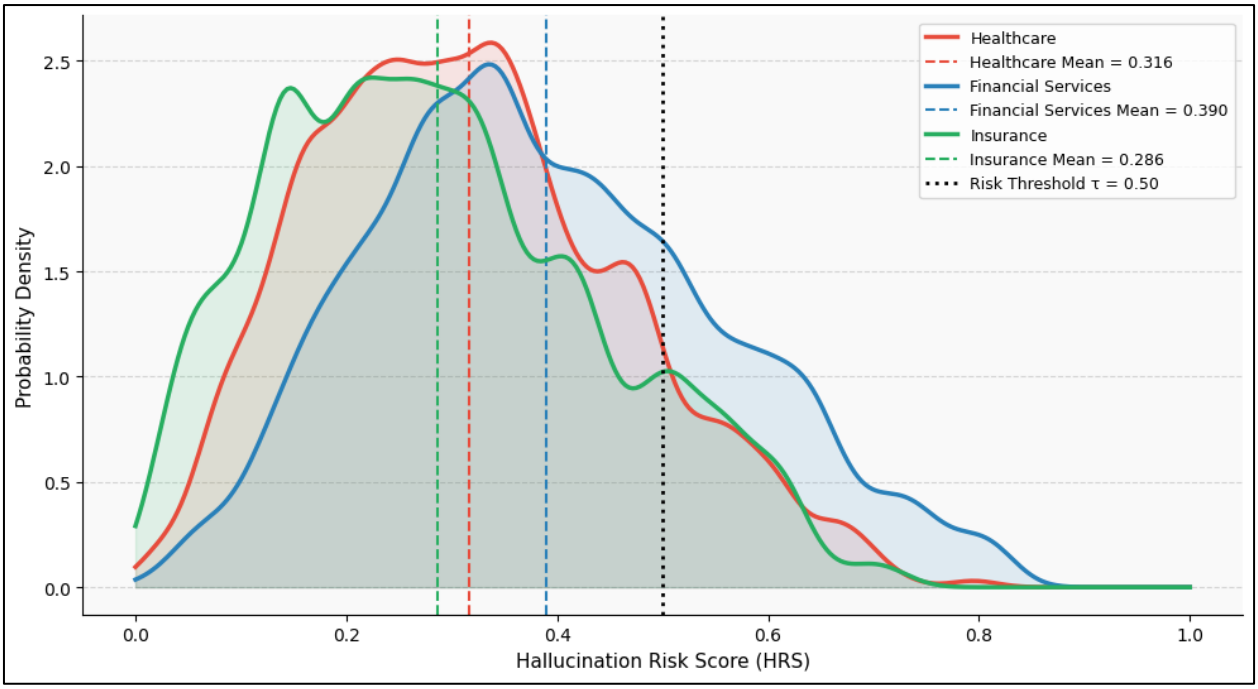

Figure 3 Hallucination Risk Score (HRS) Distribution Across Domains

This figure 3 visualizes the probability density of HRS values across all three domains, revealing the spread and skewness of hallucination risk in each sector independently of detection method performance.

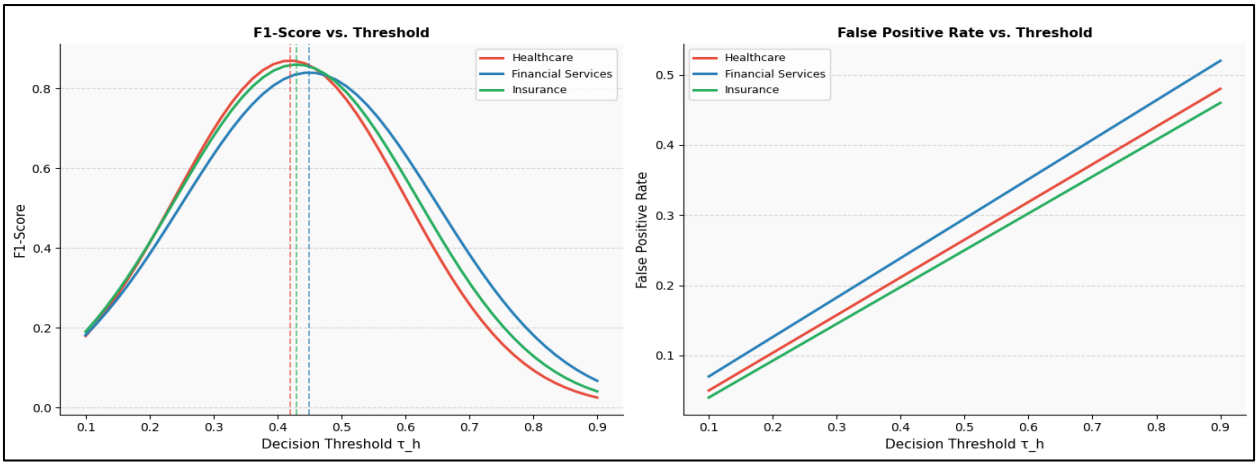

Figure 4 Threshold Sensitivity Analysis on Detection Performance

This figure 4 examines how the hallucination detection F1-score and false positive rate vary as the HRS decision threshold τh is swept from 0.1 to 0.9, providing guidance on optimal threshold selection for each domain.

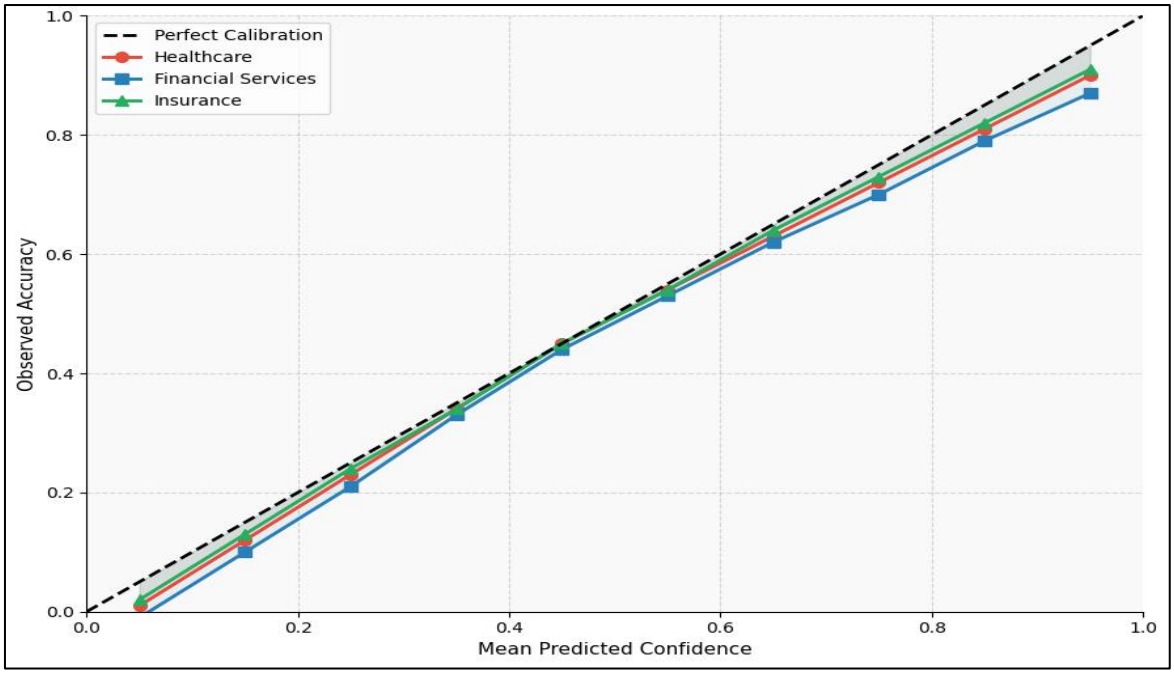

Figure 5 Model Confidence Calibration Curve

This figure 5 presents confidence calibration curves for the proposed framework across domains, comparing predicted confidence against observed accuracy to assess whether the model's uncertainty estimates are reliable for enterprise decision-making.

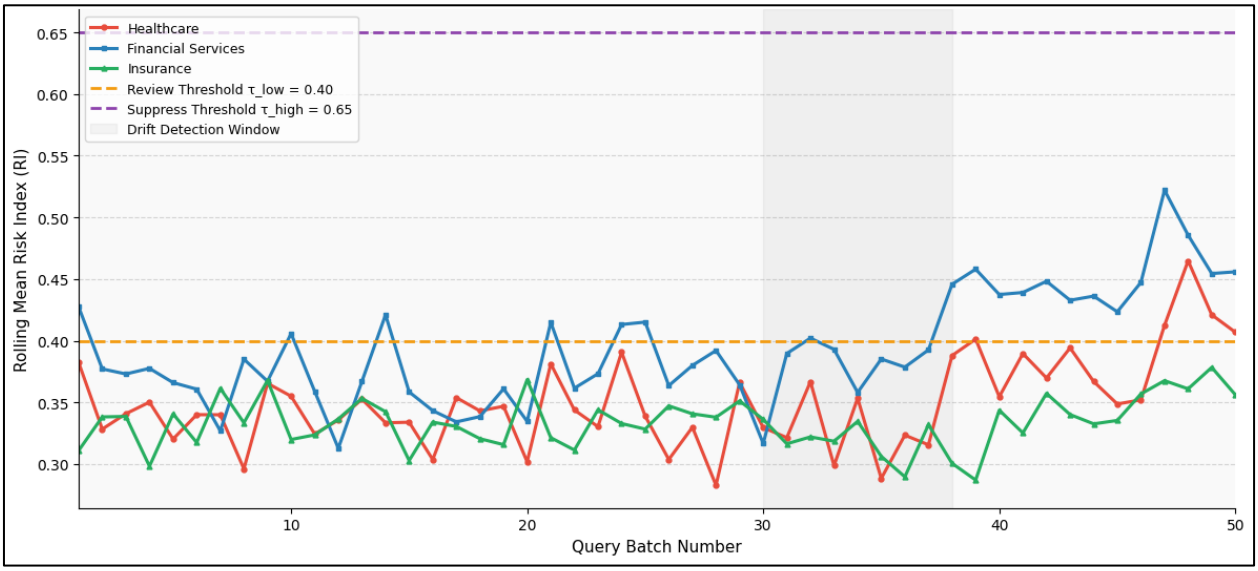

Figure 6 Risk Index Progression Over Query Batches (Temporal Monitoring)

This figure 6 tracks the rolling mean Risk Index (RI) across sequential query batches within each domain, simulating a real-world continuous monitoring scenario and demonstrating the framework's stability and drift detection capability over time.

5. Conclusion

This paper presented a comprehensive risk-centered framework for detecting, evaluating, and mitigating hallucination and bias in Large Language Models deployed within regulated enterprise environments. Unlike prior work that addresses these failure modes in isolation or under controlled benchmark conditions, the proposed framework

integrates retrieval-augmented hallucination detection, counterfactual bias evaluation, and composite risk scoring into a unified governance pipeline designed for operational enterprise deployment across healthcare, financial services, and insurance sectors.

The empirical results demonstrated that the proposed Retrieval-Augmented NLI approach achieved an average F1-score of 0.857 across all domains, outperforming baseline methods by approximately 16 percentage points. Bias mitigation yielded an average CBS reduction of 70.8% across all protected attribute dimensions, with post-mitigation scores consistently falling below the proposed enterprise fairness tolerance threshold. The temporal monitoring analysis further revealed that risk drift is a measurable and domain-specific phenomenon, reinforcing the inadequacy of static pre-deployment evaluation in regulated contexts.

Beyond performance metrics, the framework addresses the governance infrastructure requirements — audit trails, regulatory alignment, and domain-adaptive calibration — that existing approaches fail to provide comprehensively. These capabilities are not peripheral enhancements but foundational requirements for organizations operating under HIPAA, GDPR, and the EU AI Act.

Future work will explore the extension of this framework to multimodal enterprise LLMs, automated threshold recalibration under distributional shift, and the integration of explainability mechanisms to further strengthen regulatory auditability and stakeholder trust in high-stakes AI deployments.

References

[1] Gelman, H.; Hastings, J.D. Scalable and Ethical Insider Threat Detection through Data Synthesis and Analysis by LLMs. arXiv 2025, arXiv:2502.07045. [Google Scholar] [CrossRef]

[2] Portnoy, A.; Azikri, E.; Kels, S. Towards Automatic Hands-on-Keyboard Attack Detection Using LLMs in EDR Solutions. arXiv 2024, arXiv:2408.01993. [Google Scholar]

[3] Diakhame, M.L.; Diallo, C.; Mejri, M. MCM-Llama: A Fine-Tuned Large Language Model for Real-Time Threat Detection through Security Event Correlation. In Proceedings of the 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET), Sydney, Australia, 25–27 July 2024; pp. 1–6. [Google Scholar]

[4] Mudassar Yamin, M.; Hashmi, E.; Ullah, M.; Katt, B. Applications of LLMs for Generating Cyber Security Exercise Scenarios. IEEE Access 2024, 12, 143806–143822. [Google Scholar] [CrossRef]

[5] Kwan, W.C.; Zeng, X.; Jiang, Y.; Wang, Y.; Li, L.; Shang, L.; Jiang, X.; Liu, Q.; Wong, K.F. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. arXiv 2024, arXiv:2401.16745. [Google Scholar]

[6] Xu, H.; Wang, S.; Li, N.; Wang, K.; Zhao, Y.; Chen, K.; Yu, T.; Liu, Y.; Wang, H. Large language models for cyber security: A systematic literature review. arXiv 2024, arXiv:2405.04760. [Google Scholar] [CrossRef]

[7] Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv 2024, arXiv:2402.07927. [Google Scholar] [CrossRef]

[8] Chen, Y.; Cui, M.; Wang, D.; Cao, Y.; Yang, P.; Jiang, B.; Lu, Z.; Liu, B. A survey of large language models for cyber threat detection. Comput. Secur. 2024, 145, 104016.

[9] Peres, R.S.; Jia, X.; Lee, J.; Sun, K.; Colombo, A.W.; Barata, J. Industrial artificial intelligence in industry 4.0—Systematic review, challenges and outlook. IEEE Access 2020, 8, 220121–220139.

[10] Decardi-Nelson, B.; Alshehri, A.S.; Ajagekar, A.; You, F. Generative AI and process systems engineering: The next frontier. Comput. Chem. Eng. 2024, 187, 108723.

[11] Azaria, A.; Mitchell, T. The internal state of an LLM knows when it's lying. arXiv 2023, arXiv:2304.13734.

[12] Varshney, N.; Yao, W.; Zhang, H.; Chen, J.; Yu, D. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. arXiv 2023, arXiv:2307.03987.

[13] Heo, S.; Son, S.; Park, H. Halucheck: Explainable and verifiable automation for detecting hallucinations in llm responses. Expert Syst. Appl. 2025, 272, 126712.

[14] Kossen, J.; Han, J.; Razzak, M.; Schut, L.; Malik, S.; Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. arXiv 2024, arXiv:2406.15927

[15] Galitsky, B. Improving open domain content generation by text mining and alignment. In AI for Health Applications and Management; Galitsky, B., Goldberg, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2021

[16] Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. arXiv 2023, arXiv:2303.08774.

[17] Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. arXiv 2023, arXiv:2302.13971.

[18] Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. arXiv 2023, arXiv:2312.11805

[19] Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; Hashimoto, T.B. Benchmarking large language models for news summarization. Trans. Assoc. Comput. Linguist. 2024, 12, 39–57.

[20] Yu, F.; Zhang, H.; Tiwari, P.; Wang, B. Natural language reasoning, a survey. ACM Comput. Surv. 2024, 56, 1–39.