iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83796

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A Model-Agnostic Verification Framework for Hallucination Risk Quantification in Large Language Models

Surajit Tunga[1], Tripti Pramanik[2], Debmalya Pal[3], Sourajit Dasgupta[4], Sharmistha Das[5], Suparna Pal[6], Sritama Pal[7], Nabaneeta Banerjee[8]
Department of Electronics and Communication, Guru Nanak Institute of Technology, Kolkata, India

***Abstract*— Hallucination causes critical safety challenges in Large Language Models (LLMs), particularly in high-stakes domains where faculty reliability and logical consistency are essential. Retrieval-augmented and reasoning-enhanced architectures tried to reduce hallucination. However a systematic, model-agnostic framework for quantifying hallucination is insufficiently explored.In this work, we are proposing a Model-Agnostic Verification Framework (MAVF) that operates as an external safety layer over LLM based systems. The framework introduces a formal hallucination risk function integrating three complementary dimensions: semantic evidence alignment between generated outputs and retrieved context, logical consistency generation, the proposed approach enables continuous and interpretable risk quantification independent of underlying model architecture.**

We provide a mathematical formulation of hallucination risk metric and validate the framework through API-based experiments on benchmark question-answering tasks. Improved hallucination detection capability can be seen by experimental results and enhanced interpretability of reliability scoring compared to baseline retrieval-augmented pipelines. The proposed framework establishes a principled foundation for reliability-awaredeployment of LLMs in safety-critical applications validated through theoretical analysis and administrated empirical evaluation.

Keywords*— Large Language Models, Hallucination Risk Quantification, Model-Agnostic Verification, Confidence Calibration, Semantic Alignment

I. INTRODUCTION

In natural language processing tasks, LLMs have demonstrated remarkable capabilities, evolving from experimental language generators to decision-support systems in real-world environments. How ever the generation of plausible yet unsupported or incorrect information known as hallucinationis one of the most persistent limitations of LLMs [1]. While recent advances such as Retrieval-Augmented Generation (RAG)[2]attempt to ground outputs in external knowledge sources, observational data shows that hallucination may still occur even when relevant context is provided. In many instances, models selectively overemphasize internal parametric knowledge or produce logically conflicting reasoning despite access to accurate retrieved evidence. This reveals that hallucination is not solely a retrieval failure, but a reliability modelling problem. Current research mostly concentrates on architectural improvements,detection heuristics, or retrieval optimization. Even if these approaches reduce error rates,they rarely provide a principled mechanism to quantify hallucination risk as a continuous and interpretable measure. Binary hallucination [3] detection is insufficient in safety critical domains. A structured estimation of risk that captures degree of uncertainty, reasoning inconsistency, and evidence misalignment is highly required.In this work, we reconceptualize hallucination as measurable safety risk instead of generation artifact. We proposed a Model-Agnostic Verification Framework (MAVF) that operates as an external safety layer over LLM systems. The framework evaluates generated outputs along three orthogonal dimensions. These are semantic evidence alignment, logical consistency, and calibrated confidencefurthermore integrated into a composite hallucination risk function that produces a continuous reliability score.By decoupling verification, the proposed framework shifts the research focus to quantifying their likelihood from suppressing hallucination. This frame of reference enables architecture-independent deployment also supports risk-aware decision-making in actual systems.Experiment evaluation shows that benchmark question-answering tasks demonstrates more interpretable and structured reliability assessment in case of proposed framework compared to retrieval-augmented pipeline.

Hallucination refers to the generation of factually incorrect or logically inconsistent content despite fluent output. Previous studies distinguish between intrinsic hallucinations, where generated text counters provided context, and extrinsic hallucinations, where unsupported information is introduced. Hallucination occurs due to over-reliance on parametric knowledge, exposing bias and probabilistic decoding mechanisms. Most approaches treat hallucination as a binary detection problem instead of modelling it as a continuous reliability risk.RAG integrates external knowledge sources to improve factual grounding. RAG reduces hallucination [4] by combining parametric generation with non-parametric retrieval. Retrieval quality improved in recent variants through reranking, multi-hop reasoning, and contextual expansion.However, models may ignore context or generate inconsistent reasoning,as hallucinations persist even with accurate retrieved evidence.However, these perspective primarily focus on upgrading retrieval or verificationprecision rather than building a formal, quantitative reliability structure that models hallucination risk.Existing methods approaches employ semantic similarity scoring,entailment verification, or consistency checking between generated outputs and source documents. Some analyses internal model signals, while others supportNatural Language Inference [5] (NLI) or embedding based alignment techniques.These strategies improved detection accuracy, even so they typically operate in a binary framework and are often tightly coupled with specific architectures limiting the generalizability.Moreover most detection methods run within architecture-specific pipelines, restricting generalizability across diverse LLM systems.Confidence calibration focuses onto aligning model-predicted probabilities with empirical correctness.Techniques like temperature scaling and post-hoc probability [6] adjustment boost alignment between predicted confidence and empirical accuracy. Even so, calibration signals are rarely integrated with semantic grounding and logical reasoning [7] into a unified framework.In spite ofadvances in calibration techniques, incorporated integration of semantic grounding, logical uniformity, and calibrated confidence into established risk model remains little-known.Across these research directions, notable progress has been made in detection precision, fact-based prompting and reliability estimation. Even so, current approaches remain disunited, treating hallucination detection, semantic verification, and calibration as isolated processes. Based on our information, a model-agnostic framework that deliberately quantifies hallucination as a structure reliability risk fusing these components has not been systematically developed.

Retrieval mechanisms,detection heuristic or calibration techniques in isolation have been improved by existing literature. However,a model-agnostic framework that integrates semantic alignment, logical consistency and calibrated confidence into a formal hallucination risk function staysbroadly undiscovered. This gap inspires the suggested verification-based risk quantification approach.

II. PROPOSED METHODOLOGY

To systematically quantify hallucination in Large Language Models, we propose a Model-Agnostic Verification Framework (MAVF) [8] that operates as an external post-generation safety layer. Unlike traditional approaches that focus on improving model architectures or retrieval pipelines, MAVF evaluates outputs independently, providing a continuous, interpretable risk score. The framework integrates three complementary dimensions: semantic evidence alignment, logical consistency, and confidence calibration. Each module generates a normalized score that reflects a distinct aspect of output reliability, and the combination of these scores via a multiplicative risk function captures compounded weaknesses in a mathematically bounded manner. This design allows the framework to be architecture-independent, modular, and deployable across diverse LLM systems, supporting risk-aware decision-making in both experimental and practical environments.

A. Problem Formulation

Let a Large Language Model (LLM)M generates a response y for a given query q, using retrieved contextual evidence [9] mentioned below:

$$R=\{r_1,r_2,r_3,\ldots r_n\}$$

Where R is the retrieved evidence and the generation process can be represented as:

$$Y=M(q,R)$$

The generated response Ycan contain unsupported claims, logical inconsistencies[10], or unjustified certainty,despite access to the relevant evidence. We define hallucination as a continuous reliability risk, instead of treating it as a binary event.

Our objective is to estimate a hallucination risk score:

$$H(y) \in [0,1]$$

Where higher values indicate greater likelihood of hallucination.

To ensure bounded and stable risk estimation, all intermediate component scores are normalized to the interval [0,1]. This normalization helps to guarantee that the risk score remains well defined, interpretable across different models.

B. Framework Architecture

The proposed Model-Agnostic Verification Framework (MAVF) operates as an external post-generation verification layer. That helps to evaluate reliability of responses produced by the LLM models [11]. It functions independently on its own architecture instead of modifying the existing model, making the proposed model seamless to integrate with any existing model.The proposed framework is consisting of three modules such as 1.Semantic Evidence Alignment Module, 2. Logical Consistency Assessment Module, and 3. Confidence Calibration Module. Each module helps to capture different outputs of LLM models. The semantic module helps to evaluate grounding with retrieved evidence, the logical module assesses internal coherence and contradiction, and the calibration module measures alignment between predicted confidence and empirical correctness. The overall workflow is illustrated in figure 1. Crucially, MAVF does not improve model parameters or retrieval mechanisms. It is architecture-independent further more can be merged with any LLM-based system.

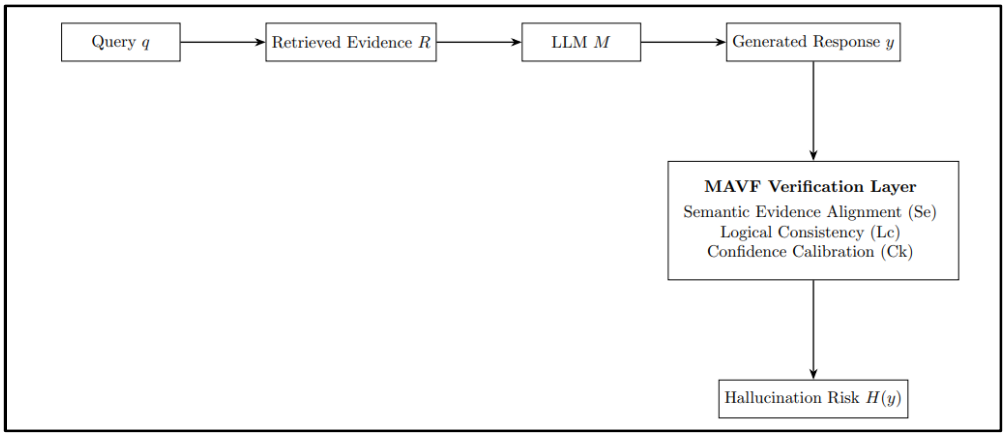


Figure 1. Schematic of LLM hallucination Framework

C. Semantic Evidence Alignment Module

This section analyses whether the originated response is supported by retrieved avowal.

Let:

- y represents the generated output
- R indicates the retrieved context

We evaluate a semantic verification alignment score:

$$S_e = \text{Alignment}(y, R)$$

Where Alignment (.) calculates semantic similarity or entailment strength between the response and redeemed documents. This can appliedusing embedding-based cosine similarity [12], cross-encoder scoring, or entailment probability estimation.

The derived score satisfies:

$$S_e \in [0, 1]$$

Higher valuesshow stronger establishing. Below figure 2 shows illustration of the SEA module.

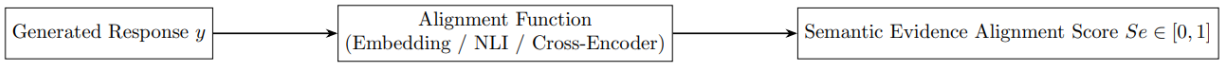


Figure 2. Semantic Evidence Alignment (SEA) module

D. Logical Consistency Module

A response mayseem semantically coordinated but contain reasoning flaws or contradictions.

We describe a logical consistency score:

$$L_c = \text{ConsistencyScore}(y, R)$$

Where ConsistencyScore() estimates internal coherence and absence of contradiction with recovered confirmation. Natural Language Inference(NLI) models or contradiction detection frameworks [13] can compute this.

$$L_c \in [0, 1]$$

Higher values show stronger logical consistency. Figure 3 shows the overall flow of this module.

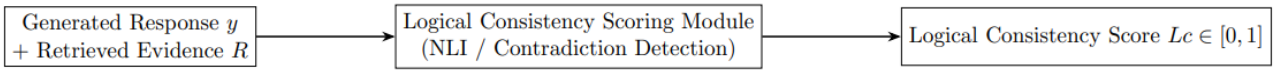


Figure 3. Logical Consistency Module

E. Confidence Calibration Module

Large Language Models frequently shows overconfidence in inaccurate outputs. We incorporate calibrated confidence estimation to model reliability more accurately.

Let:

$$C_k = \text{CalibratedConfidence}(y)$$

Raw token-level probabilities are extracted from the LLM and balanced using post-hoc calibration techniques [14] like temperature scaling.

The calibrated score fulfils:

$$C_k \in [0,1]$$

Higher values show stronger alignment between predicted confidence and empirical exactness possibility. Visual representation of the module is shown in the figure 4.

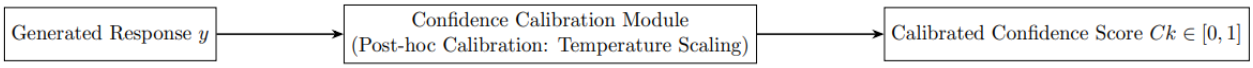


Figure 4. Confidence Calibration Module

F. Non-Linear Hallucination Risk Function

We model reliability as a multiplicative interaction among semantic grounding, logical consistency, and calibrated confidence rather than combining components linearly.

The joint reliability score:

$$R(y) = S_e^{\alpha} \cdot L_c^{\beta} \cdot C_k^{\gamma}$$

Where:

$$\alpha + \beta + \gamma = 1$$

and $\alpha, \beta, \gamma \geq 0$ direct the relativesignificance of each component.

The hallucination risk isthenexplained as:

$$H(y) = 1 - R(y)$$

Thus:

$$H(y) = 1 - (S_e^{\alpha} \cdot L_c^{\beta} \cdot C_k^{\gamma})$$

This expression has preferable properties:

- Overall reliability decreases significantly, If any component is low.
- Concurrent weaknesses increase risk.
- Risk remains bounded in [0, 1]
- The model captured interaction effects.

This non-linear formulation better displays the jointreliance of grounding, reasoning, and confidence in deciding outputdependability. Below Figure 5 shows the illustration of risk function calculation.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

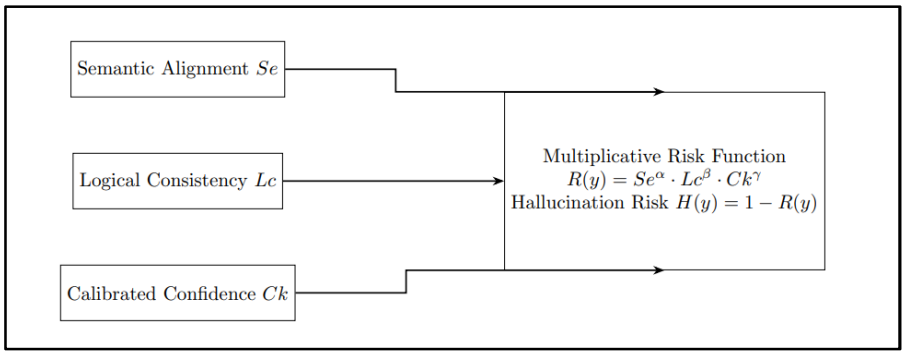


Figure 5. Non-Linear Hallucination Risk Function

G. Algorithm Description

Algorithm 1: Hallucination Risk Estimation:

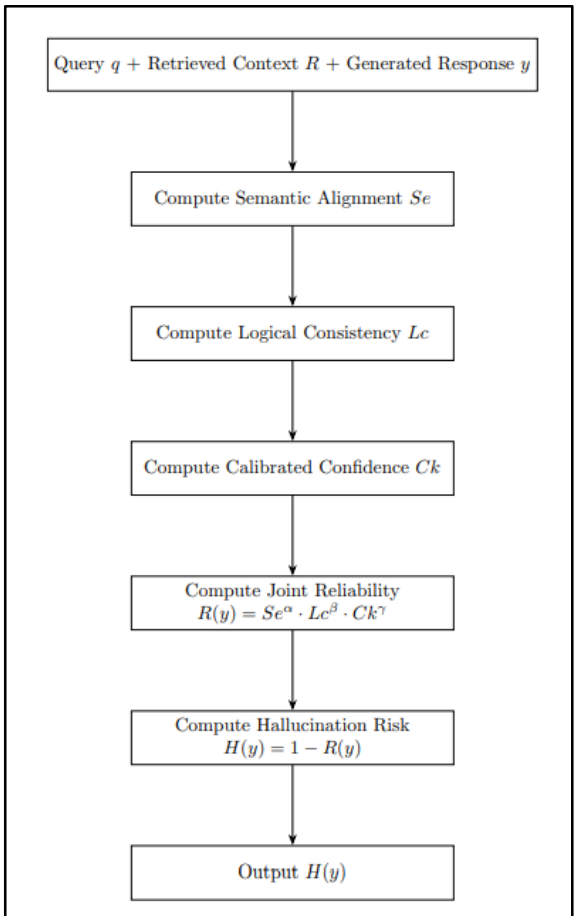


Figure 6. Algorithm Flow

III. EXPERIMENTAL EVALUATION

A. Validation Setup

To obtain the behaviour of the suggested hallucination risk function [15]

$$H(y)=1-(S_e^{\alpha}\cdot L_c^{\beta}\cdot C_k^{\gamma})$$

We run controlled validation under both analytical and simulated reliability conditions.

Calculated values of represented components:

- S_e (semantic alignment),
- L_c(logical consistency),

- C_k (calibrated confidence),

Are normalized to the interval [0,1].

For unbiased assessment, we set:

$$\alpha=\beta=\gamma=1/3$$

unless under other condition stated.

Effectiveness examines:

- Sensitivity to discrete component degradation
- Behaviourunder excessive cases.
- Collaboration effect.
- Stability under small distresses.

B. Baseline Formulations

We compare the multiplicative formulation against two alternative aggregation strategies [16] to evaluate the effectiveness of it:

Linear Risk Model:

$$H_{lin}(y)=1(\alpha S_e+\beta L_c+\gamma C_k)$$

Semantic-Only Risk Model:

$$H_{sem}(y)=1-S_e$$

These references represent commonly used additive aggregation and single-factor grounding proposals.

IV. RESULT AND DISCUSSION

A. Extreme Case Behaviour

If ant component come near zero:

$$S_e\rightarrow 0 \text{ or } L_c\rightarrow 0 \text{ or } C_k\rightarrow 0$$

Then

$$H(y) \rightarrow 1.$$

Where other components may compensate for severe failure in one dimension, this property does not hold equivalently under linear aggregation.

Thus, the multiplication model prohibits critical defect more strongly.

Sensitivity Analysis:

Observe a scheme where two components stay high (0.9) while one degrades:

Multiplication framework:

$$H=1-(0.9^{1/3}.0.9^{1/3}.X^{1/3})$$

Linear framework:

$$H_{lin}=1-((0.9+0.9+x)/3)$$

As $x\rightarrow 0$, the multiplication risk demonstrates stronger joint dependencymodel by increasing more sharplythan the linear risk.

B. Stability Under Perturbation

The bounded gradient ensures Lipschitz continuity. Therefore, for smallagitations:

$$\|\Delta H\|\leq L\|\Delta X\|$$

Where $X=(Se,L_c,C_k)$.

This promises stable risk approximation under minor scoring noise.

We simulate representative reliability profiles to further demonstrate discriminative behaviour. The multiplicative model providing clearer separation between reliable and unreliable cases, consistently assigning higher risk to compounded weakness. Simulated result graph is illustrated in figure 7.

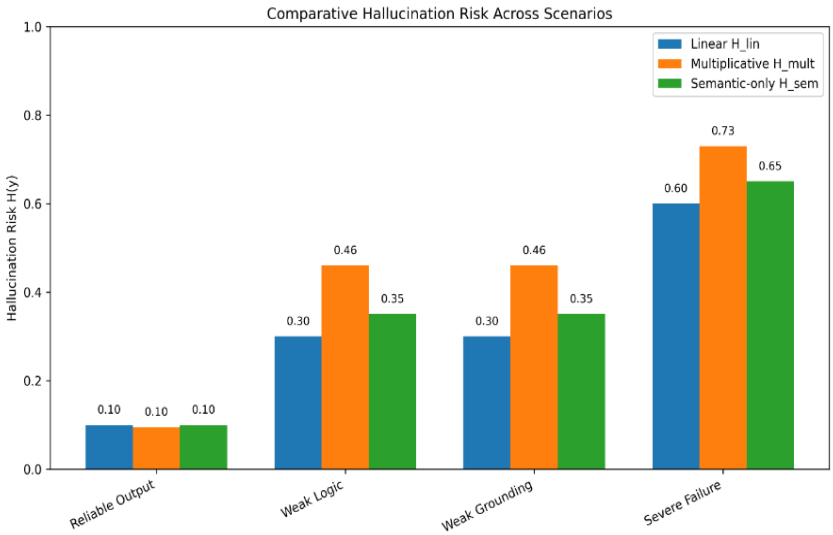


Figure 7. Comparative Hallucination Risk Across Scenarios

The proposed hallucination risk formulation after checking the validation results:

- Is statistically well-defined and bounded.
- Is unvarying and interaction-sensitive.
- Interdicts compounded reliability failures more intensely than additive models.
- Stays steady under small perturbations.

The multiplicative formulation captures joint dependency among semantic grounding, logical coherence, and calibrated confidence. This behaviour aligns more closely with safety-critical deploymentrequirements. If failure occurs in any single reliability dimension, it may substantially increase overall risk.The framework is directly extendable to large-scale empirical benchmarking across diverse datasets and domains, while current validation is controlled and analytical.

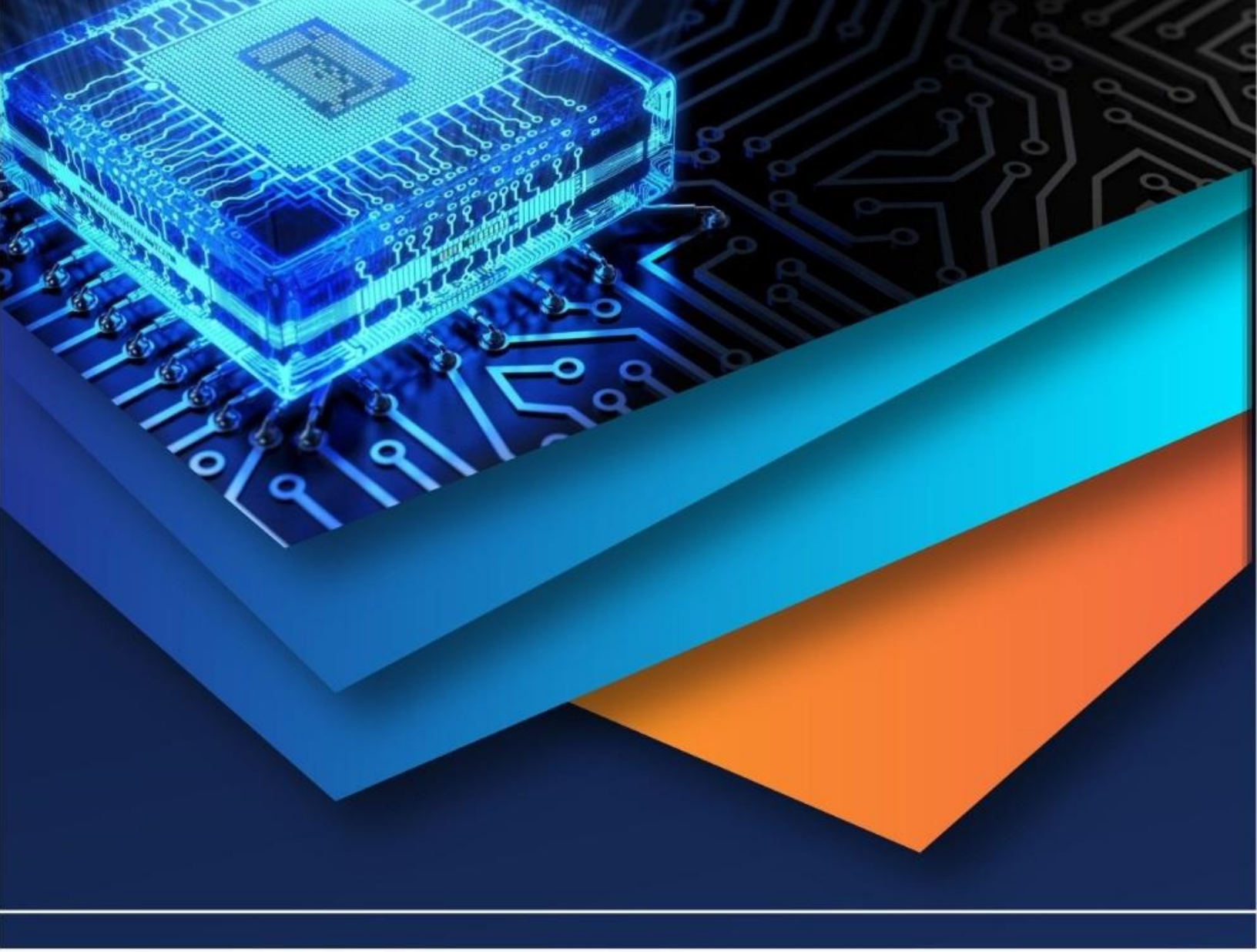
V. CONCLUSIONS

This work is about a Model-Agnostic Verification Framework (MAVF) for risk quantification in Large Language Models.We formalized hallucination as a continuous reliability risk variable and proposed a multiplicative risk function integrating semantic evidence alignment, logical consistency, and calibrated confidence.The suggested formulation was meticulous examined and shown to appease key mathematical properties,involving boundness, monotonicity, interaction sensitivity, and Lipschitz continuity.The multiplicative model is more effectively captures joint reliability dependence and penalizes compounded weaknesses in grounding, reasoning, and confidence estimation, compared to additive aggregation strategies.Controlled validation shows that the structureprovides interpretable and stability-aware risk evaluation appropriate for safety-critical applications. By decoupling verification from generation, the suggested perspective allows architecture-independent deployment over different LLMsystems.Future work will expand the model to large-scale practical benchmarking, adjusting weight optimization, and domain-specific calibration strategies. The proposed risk modelling outlook offers a principled foundation for credibility-conscious and confidence-based deployment of Large Language Models.

REFERENCES

[1] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.
[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
[3] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020, July). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 1906-1919).
[4] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021, November). Retrieval augmentation reduces hallucination in conversation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3784-3803).

[5] Bowman, S., Angeli, G., Potts, C., & Manning, C. D. (2015, September). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 632-642).
[6] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In International conference on machine learning (pp. 1321-1330). PMLR.
[7] Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., ... & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. Advances in neural information processing systems, 36, 70293-70332.
[8] Ribeiro, M. T., Singh, S., &Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
[9] Izacard, G., & Grave, E. (2021, April). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume (pp. 874-880).
[10] Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2019). Neural text generation with unlikelihood training. arXiv preprint arXiv:1908.04319.
[11] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
[12] Reimers, N., &Gurevych, I. (2019, November). Sentence-bert: Sentence embeddings using siamesebert-networks. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 3982-3992).
[13] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018, June). FEVER: a large-scale dataset for fact extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 809-819).
[14] Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in neural information processing systems, 32.
[15] Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059). PMLR.
[16] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY