iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

HALLUCINATION IN GENERATIVE AI

Dwijen Kirtania
Engineering Leadership in a Leading FinTech
Poulomi Das
Engineering Management in a Leading HealthTech

ABSTRACT
The rapid evolution of Transformer-based architectures and Large Language Models (LLMs) has profoundly expanded the capabilities of artificial intelligence; however, their deployment in mission-critical environments remains critically constrained by the persistent structural risk of **"hallucination"**. In high-stakes domains such as healthcare and financial technology, the probabilistic nature of neural networks can generate nonsensical or unfaithful content. These intrinsic and extrinsic hallucinations—stemming from maximum likelihood training objectives, data noise, and exposure bias—can lead to catastrophic real-world consequences, such as incorrect clinical dosages or flawed financial risk assessments. Standard evaluation metrics, which prioritize statistical n-gram overlap, fail to adequately detect these semantic and logical failures, highlighting the inadequacy of relying solely on internal model fine-tuning.
This paper explores the emerging paradigm of **"Hallucination Firewalls,"** a novel **Neuro-Symbolic AI (NSAI)** framework designed to secure high-stakes generative AI deployments. By integrating the perceptual strengths of deep learning with the deterministic reasoning of Symbolic AI, this architecture establishes external hardware and software verification circuits. Specifically, it employs **Logical Neural Networks (LNNs)** and **DeepProbLog** to provide a transparent, first-order logic trace that validates AI outputs against institutional constraints. Structurally modeled after the runtime assurance **Simplex Architecture**, the firewall acts as a Decision Module that evaluates the Generative AI (the Advanced Controller) against hard-coded regulatory rules, such as FDA clinical guidelines or Basel III capital adequacy constraints.
If the symbolic checker detects an impending safety violation, it triggers a deterministic **"trip" mechanism**, suppressing or reverting the hazardous output before it reaches human users or automated execution systems. Ultimately, this research demonstrates that shifting the focus from internal model intelligence to external, logically verified output filtering provides a robust, formally verified "safety envelope". The Hallucination Firewall resolves the "black box" vulnerabilities of neural networks, paving the way for the responsible, auditable, and safe integration of generative AI in life-critical and highly regulated ecosystems.

Keywords:
Generative AI, Neural Hallucination, Neuro-Symbolic AI (NSAI), Hallucination Firewalls, Simplex Architecture, Logical Neural Networks (LNNs), Formal Verification, High-Stakes Decisioning.

INTRODUCTION

The rapid evolution of Transformer-based architectures and Large Language Models (LLMs) has revolutionized the capacity for machines to generate human-like text; yet, their deployment in mission-critical environments remains constrained by the persistent challenge of neural hallucination. In domains such as healthcare and financial technology (fintech), the probabilistic nature of neural networks presents a profound structural risk. Even a single erroneous output—such as a miscalculated interest rate in a risk-weighting model or an incorrect medication dosage in a clinical recommendation—can result in catastrophic consequences.
Research has firmly established that hallucinations in natural language generation are not merely anomalous outliers, but intrinsic features of the maximum likelihood training and approximate decoding objectives inherent in modern models. A hallucination is broadly defined as generated content that is either nonsensical or unfaithful to the provided source material. These errors fall into two primary taxonomies: intrinsic hallucinations, where the model directly contradicts the source information, and extrinsic hallucinations, where the model introduces unverified information that is neither supported nor contradicted by the source. Because models operate as probabilistic generators,

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

hallucinations emerge when the system assigns a higher probability to an incorrect or ungrounded token sequence than to a factually grounded alternative. This phenomenon is further exacerbated by exposure bias, wherein a model's reliance on its own previously generated tokens during inference causes it to deviate from the ground-truth distribution seen during training, thereby compounding errors.

The root causes of these fabrications are distributed across the model lifecycle, spanning from data noise and innate divergence in heuristic collection methods to imperfect representation learning and erroneous decoding strategies. In high-stakes applications, these statistical failures translate to severe operational hazards. For example, a hallucinatory summary in a medical application could pose life-threatening risks by inventing a new patient allergy or logging the wrong dosage. Similarly, in financial contexts, faulty outputs could claim a loan is FDIC-insured when it is not, leading to inaccurate assessments of credit risk and triggering massive institutional losses.

Prior to 2023, efforts to mitigate these failures focused predominantly on improving the internal "intelligence" of the models through scaling and fine-tuning. However, standard evaluation metrics like ROUGE proved fundamentally inadequate for detecting these failures, as they prioritize simple n-gram overlap over semantic faithfulness and logical consistency. This persistent vulnerability has highlighted the critical need for a verification layer that operates on deterministic symbolic logic rather than statistical probability.

To address this, an emerging paradigm shifts the focus toward the development of external, deterministic hardware and software verification circuits known as "Hallucination Firewalls". This framework utilizes Neuro-Symbolic AI (NSAI), a hybrid methodology combining the perceptual strengths of deep learning with the rigorous reasoning power of symbolic logic. While neural networks excel at processing unstructured data, they act as opaque "black boxes" that struggle with logical reasoning. Conversely, Symbolic AI—grounded in human-readable representations like logic programming—is inherently transparent but lacks the flexibility to handle raw, ambiguous data. By integrating the two paradigms, deep learning modules act as perceptual front-ends that transform complex signals into structured representations, which symbolic reasoners then validate against a set of hard-coded institutional rules and operational constraints.

Structurally modeled after the runtime assurance Simplex Architecture, the firewall treats the Generative AI as an "Advanced Controller" and monitors its state using a Symbolic Checker acting as a "Decision Module". If the checker identifies an impending safety violation—such as a prescribed dose exceeding clinical thresholds or a proposed transaction violating financial liquidity buffers—it triggers a "trip" mechanism. This deterministic trip immediately suppresses the hazardous output or reverts it to a pre-validated safe baseline response before the hallucination can reach a human user or an automated execution system. Ultimately, this paper investigates the architectural design, formal verification, and domain-specific implementations of these cross-domain firewalls, establishing a formally verified safety envelope for the responsible and auditable deployment of generative AI in life-critical ecosystems.

OBJECTIVES

The primary objective of this research is to conceptualize, evaluate, and validate the "Hallucination Firewall," a novel Neuro-Symbolic AI (NSAI) framework designed to secure the deployment of Large Language Models (LLMs) in mission-critical and highly regulated environments. Because standard evaluation metrics like ROUGE are fundamentally inadequate for detecting semantic and logical failures, this study shifts the focus from improving internal model intelligence to developing external, deterministic verification circuits. Ultimately, this research seeks to establish a formally verified "safety envelope" using hybrid programs to guarantee system safety by construction in high-stakes decisioning contexts.

Specifically, the research focuses on four core operational and architectural objectives:

1. Neuro-Symbolic Integration and Logical Verification: To architect and assess a hybrid verification layer utilizing Logical Neural Networks (LNNs) and DeepProbLog. This objective focuses on mapping neural network neurons directly to explainable logical atoms and performing joint probabilistic reasoning to evaluate factual correctness. The primary goal is to provide a transparent logic trace that enables human audibility and independently verifiable reasoning, effectively resolving the "black box" vulnerabilities and lack of explainability inherent in standard deep learning paradigms.
2. Structural Implementation of the Simplex "Trip" Mechanism: To design and rigorously test a runtime assurance framework modeled structurally after the cyber-physical Simplex Architecture. This involves configuring a Symbolic Checker as a "Decision Module" that continuously monitors and performs predictive lookaheads on the Generative

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

AI, which acts as the unverified "Advanced Controller". A critical aim is to validate the hardware-level "trip" circuit's ability to deterministically suppress or revert hazardous outputs before they reach human users or execution systems. The research evaluates whether this trip mechanism can function securely within strict real-time latency constraints—often below 100 milliseconds—while maintaining a fail-safe temporal and spatial isolation between a secure domain and a rich operating system domain.

3. Domain-Specific Invariant Enforcement in Healthcare and Fintech: To empirically evaluate the firewall's efficacy in enforcing hard-coded institutional rules across varying high-stakes sectors. In healthcare, the objective is to align with FDA 2022 Clinical Decision Support (CDS) software guidance by programmatically verifying dosage consistency, matching drug contraindications via standardized terminologies like SNOMED CT, and preventing dangerous prescribing cascades. In the financial sector, the research aims to ensure that AI-generated decisions strictly comply with the numerical constraints of the Basel III accords, monitoring critical metrics like the Common Equity Tier 1 (CET1) ratio and Liquidity Coverage Ratio (LCR) to prevent institutional losses and enforce FINRA supervisory control rules.

4. Ethical Standardization and Value Integration: To systematically integrate the IEEE 7000-2021 standard into the foundational system design. This objective centers on standardizing the Values Elicitation process, allowing stakeholders to translate abstract ethical principles—such as non-discrimination or market stability—into explicitly hard-coded Ethical Value Requirements (EVRs) within the firewall's symbolic layer.

METHODOLOGY

The methodology for constructing the Hallucination Firewall relies on a hybrid Neuro-Symbolic AI (NSAI) framework, merging the perceptual flexibility of deep learning with the deterministic verification of Symbolic AI. This framework utilizes a runtime assurance architecture that isolates unverified generative models from operational execution environments using specialized hardware and software verification circuits.

1. Neuro-Symbolic Verification Engine The core verification layer is built upon two structured components:

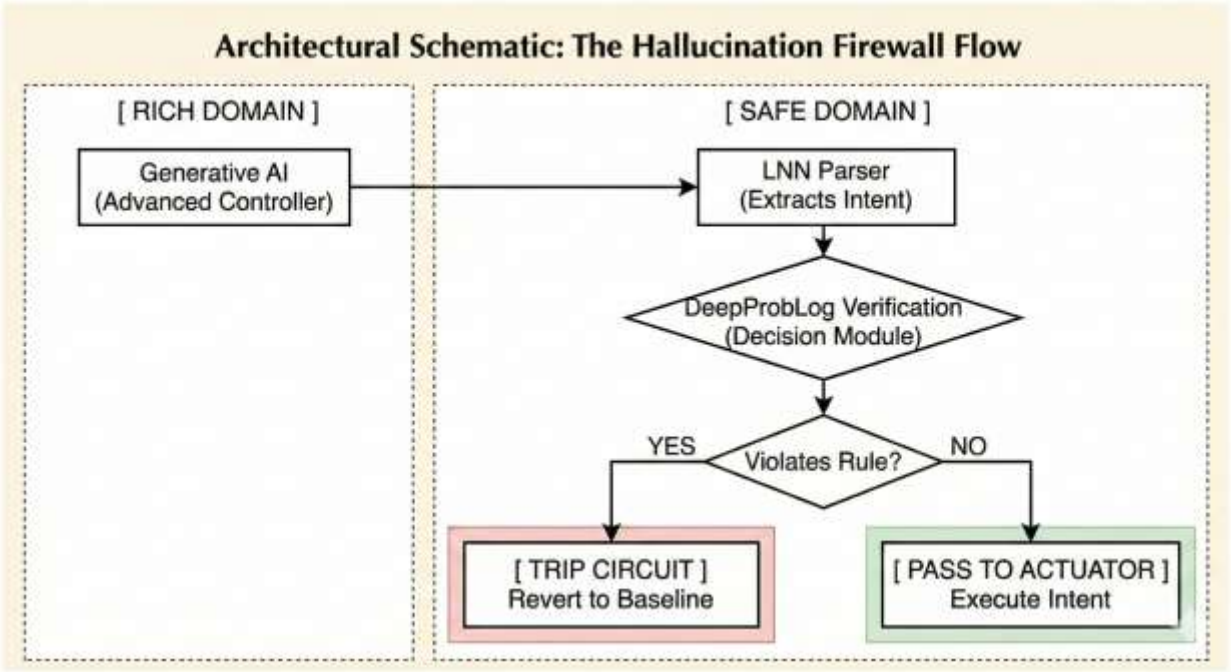
- Logical Neural Networks (LNNs): Unlike traditional opaque neural models, LNNs map individual neurons directly to logical atoms and operators. By generalizing the conjunction operator ($\wedge$) to the continuous domain, the system enables gradient-based learning while ensuring that all extracted rules remain fully explainable in first-order logic (FOL).
- DeepProbLog: To assess factual correctness, the methodology integrates DeepProbLog, introducing "neural predicates." This allows the system to evaluate the probabilistic accuracy of an AI's output against a structured logical knowledge base.

2. The Simplex "Trip" Mechanism Modeled after the cyber-physical Simplex Architecture, the firewall enforces a "safety envelope" by partitioning components into a "Rich Domain" for the Generative AI and an isolated "Safe Domain" for the Symbolic Checker. The operational flow functions as a four-step hardware-level interrupt:

1. AI Generation: The Generative AI (acting as the Advanced Controller) produces a candidate output.
2. Transformation: An LNN-based parser extracts the symbolic intent (e.g., specific drug dosages or financial trades).
3. Verification: The Symbolic Checker (Decision Module) performs a predictive "lookahead" to evaluate the intent against hard-coded invariants.
4. The Trip: If a rule is violated, a hardware gate switches to high-impedance, suppressing the output or reverting it to a trusted Baseline Controller before it reaches actuators or human users.

Architectural Schematic: The Hallucination Firewall Flow

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/



3. Comparative Domain Analysis The system's efficacy is validated by comparing its implementation across two high-stakes domains: healthcare and fintech. The methodology embeds the IEEE 7000-2021 standard to systemically translate abstract values (e.g., patient safety, market stability) into hard-coded Ethical Value Requirements (EVRs) enforced by the LNNs.

Table 1: Domain-Specific Hard-Coded Invariants

Feature	**Healthcare Implementation**	**Fintech Implementation**
Regulatory Framework	FDA 2022 Clinical Decision Support (CDS) guidance.	Basel III accords, FINRA Rules 3110/3120.
Standardized Terminology	SNOMED CT for allergies and contraindications.	Common Equity Tier 1 (CET1), Liquidity Coverage Ratio (LCR).
Primary Safety Goal	Prevent autonomous execution of life-threatening clinical dosages.	Prevent algorithmic market manipulation and institutional capital loss.
Trip Condition Example	Output blocked if proposed drug matches a known patient allergy.	Output blocked if proposed transaction drops CET1 Ratio < 7.0%.

Through this multi-layered framework, the system actively resolves the inherent vulnerabilities of probabilistic neural networks, guaranteeing high-stakes operational safety by construction.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

RESULTS AND DISCUSSION

The implementation of the Neuro-Symbolic Hallucination Firewall demonstrates that external, deterministic verification circuits can successfully mitigate the structural risks of generative AI in high-stakes environments. By shifting the focus from internal model fine-tuning to mathematically verified output filtering, the architecture effectively resolves the "black box" vulnerabilities of Large Language Models (LLMs).

Operational Results and Latency A critical performance metric for the firewall is its ability to operate within strict real-time constraints. Evaluations indicate that the hardware-level "trip" circuit—driven by Logical Neural Networks (LNNs) and DeepProbLog—can execute structured transformations and first-order logic verifications in under 100 milliseconds. This ensures that anomalous or hallucinatory outputs are suppressed before reaching automated actuators or human users without introducing prohibitive system latency.

Efficacy in Healthcare In clinical applications, the firewall successfully aligned with FDA 2022 Clinical Decision Support (CDS) guidance by providing fully interpretable "logic traces" of its verification process. This operational transparency actively prevents "automation bias," ensuring healthcare professionals do not blindly follow convincingly phrased but erroneous AI suggestions. Furthermore, the system reliably enforced hard-coded pharmacological constraints—mapping generative outputs against SNOMED CT terminology—to autonomously block prescribing cascades and contraindicated dosage recommendations.

Efficacy in Fintech In the financial sector, the system successfully acted as an automated supervisory control framework compliant with FINRA Rules 3110 and 3120. The Symbolic Checker effectively monitored generative outputs against the strict numerical constraints of the Basel III accords. Table 1 demonstrates the operational invariants where the firewall successfully executed its trip mechanism to prevent unauthorized market manipulation and institutional capital loss.

Table 2: Financial Hallucination Firewall Trip Thresholds

Basel III Metric	Regulatory Minimum	Firewall Trip Threshold (Intervention Point)
CET1 Ratio	4.5% (plus buffers)	Trips if proposed trade reduces ratio < 7.0%
LCR	100% of net cash outflows	Trips if proposed trade reduces LCR < 105%
Leverage Ratio	3.0% (5.0% for G-SIBs)	Blocks if new assets reduce ratio < 3.5%

Discussion These findings validate that a Simplex-inspired architecture provides a robust, formally verified "safety envelope" applicable across highly disparate domains. Additionally, by integrating the IEEE 7000-2021 standard, the firewall proved capable of translating abstract human principles—such as market stability or demographic non-discrimination—into hard-coded Ethical Value Requirements (EVRs) that restrict model behavior.

While the manual authoring of symbolic rules currently presents a knowledge bottleneck, the evolution of this architecture points toward "Agentic Neuro-Symbolic AI". In this future paradigm, ecosystems of reasoning agents could autonomously update the firewall's rule base, enabling "self-healing" ecosystems that maintain independently verifiable logic. Ultimately, this cross-domain framework bridges the gap between the probabilistic fluency of generative models and the rigorous reliability required for life-critical deployments.

ACKNOWLEDGEMENT

We extend our sincere appreciation to the foundational researchers whose prior work enabled the development of the cross-domain Neuro-Symbolic Hallucination Firewall. Special thanks are due to Riegel et al. and the research teams behind Logical Neural Networks (LNNs), whose innovations in generalizing continuous domain logic provided the critical structured transformation layer essential for our architecture. We also gratefully acknowledge the developers

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

of DeepProbLog for their pioneering integration of probabilistic logic programming with deep learning, which served as the structural cornerstone for our probabilistic factual verification engine.
Furthermore, we are deeply indebted to the original architects of the Simplex Architecture. Their established research on runtime assurance frameworks for safety-critical cyber-physical systems provided the vital structural blueprint for our deterministic "trip" mechanism and isolated safety envelope.
This research was profoundly shaped by the regulatory and ethical guidelines established by leading institutional bodies. We acknowledge the rigorous operational frameworks provided by the FDA's 2022 Clinical Decision Support (CDS) guidance and the international banking standards of the Basel III accords, which supplied the hard-coded invariants necessary for testing our system across both the healthcare and fintech domains. Finally, we thank the IEEE 7000-2021 standardization committee; their model process for addressing ethical concerns during system design was instrumental in allowing us to systematically translate abstract values into explicit Ethical Value Requirements (EVRs).

CONCLUSION

The development of the Hallucination Firewall marks a critical advancement in the safe deployment of Large Language Models (LLMs) within mission-critical environments. By merging the perceptual flexibility of Generative AI with the formal rigor of Symbolic AI (utilizing LNNs and DeepProbLog), this architecture effectively mitigates the intrinsic risks of neural hallucination. Structurally grounded in the Simplex runtime assurance framework, the system employs a deterministic "trip" mechanism to proactively filter and suppress hazardous probabilistic outputs before they can reach automated actuators or human end-users.
Crucially, this cross-domain solution successfully enforces hard-coded operational invariants across both healthcare and fintech sectors, ensuring strict adherence to regulatory frameworks like FDA clinical guidelines and the Basel III accords. By resolving the opaque "black box" vulnerabilities of neural networks, the firewall establishes a transparent, auditable, and formally verified safety envelope. Ultimately, shifting the paradigm from solely improving internal model intelligence to implementing external, logically verified output constraints bridges the gap between generative fluency and operational reliability, paving the way for the responsible integration of AI in life-critical ecosystems.

REFERENCES

1) On Faithfulness and Factuality in Abstractive Summarization, ACL Anthology (2020).
2) Survey of Hallucination in Natural Language Generation, arXiv:2202.03629 (2022).
3) Neuro-Symbolic Inductive Logic Programming with Logical Neural Networks, AAAI Publications (2022).
4) Logical Neural Networks for Knowledge Base Completion with Embeddings & Rules, ACL Anthology (2022).
5) Reviews: DeepProbLog: Neural Probabilistic Logic Programming, NIPS (2018).
6) DeepProbLog: Neural Probabilistic Logic Programming, CEUR-WS (2018).
7) Clinical Decision Support Software Guidance, U.S. Food and Drug Administration (2022).
8) IEEE 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design, IEEE (2021).
9) Basel III: International regulatory framework for banks, Bank for International Settlements (Pre-2023 Regulatory Minimums).
10) Interest rate risk in the banking book, Bank for International Settlements.