International Journal of Artificial Intelligence
p-ISSN: 2407-7275, e-ISSN: 2686-3251

Original Research Paper

RAG-Guardrails Integration for AI Content Control

Rakesh More[1*]

[1] *NVIDIA. Schaumburg IL, USA.*

Article History
Received:
19.07.2025

Revised:
08.08.2025

Accepted:
21.08.2025

***Corresponding Author:**
Rakesh More
Email
rakeshmore@gmail.com

Abstract: Generative AI is particularly Large Language Models (LLMs), has shown remarkable potential across domains such as healthcare, legal services, and finance. However, their adoption is hindered by two persistent challenges: hallucination, where models generate factually incorrect information and the risk of producing biased or unsafe content. This paper proposes a hybrid framework that integrates Retrieval-Augmented Generation (RAG) with NVIDIA NeMo Guardrails to address these concerns. RAG mitigates hallucinations by grounding model outputs in externally retrieved, trusted data sources, while NeMo Guardrails enforce domain-specific safety and compliance constraints through predefined behavioral policies. Empirical evaluations demonstrate that this combined approach reduces hallucinated content by 30–45% and improves safety and policy adherence across multiple enterprise use cases. The system exhibits strong potential for deployment in regulated, high-stakes environments. Future work will focus on enhancing real-time responsiveness and expanding multilingual and culturally adaptive capabilities. The proposed framework offers a scalable foundation for building trustworthy, domain-aligned generative AI solutions.

Keywords: Artificial Intelligence, Healthcare Access, Healthcare Challenges, Healthcare Delivery, Telemedicine.

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

1. Introduction

Large language models, like GPT-4 from OpenAI, Google's PaLM, and Meta's LLaMA—have really shaken up how we approach language tasks on computers. Built on top of enormous collections of text, these models use complicated architecture (transformers, mostly) to make eerily fluent language predictions. As a result, we're relying on them more than ever, whether it's powering virtual assistants, helping automate business summaries, delivering tutoring experiences, or answering random questions with speed and confidence. Benchmark studies consistently show that they excel at generating text [1], handling translation [2] and summarizing documents [3].

Despite all the hype and impressive stats, a persistent issue remains hallucinations. It's not a science fiction problem, just a term for when the model confidently "makes stuff up" that sounds convincing but isn't grounded or sometimes can't be verified at all [4]. This quickly goes from quirky to potentially dangerous if you imagine an LLM writing up a medical answer that's wrong [5] or producing a news summary that misleads readers [6]. Even in research, users have caught models fabricating references that look real but aren't tied to any actual publication [7].

Another minefield: unsafe or non-compliant outputs. Without guardrails, these language models can blurt out responses that are biased, offensive, or cross legal boundaries. This poses a major challenge for deploying these tools in fields like healthcare, law, or education, where accuracy and sensitivity are essential [8].

To address these risks, I propose a straightforward yet promising hybrid strategy. This approach integrates RAG with NVIDIA NeMo Guardrails to create a comprehensive solution addressing both hallucination and safety concerns. Our approach leverages the complementary strengths of both methodologies: RAG provides external knowledge grounding to reduce hallucinations, while NeMo Guardrails ensures comprehensive safety enforcement across all interaction phases [9] [10].

The key contributions of this work include:

- Integration Architecture
  A systematic framework for combining RAG and NeMo Guardrails that addresses the limitations of individual approaches while maximizing their complementary benefits.
- Comprehensive Evaluation
  Systematic assessment across multiple domains (healthcare, legal, finance) using established benchmarks and novel evaluation metrics specifically designed for hybrid systems.
- Security and Privacy Analysis
  Detailed examination of security implications, privacy protection mechanisms, and compliance considerations for enterprise deployment.
- Deployment Guidelines
  Practical recommendations for implementing the integrated framework in production environments, including performance optimization and scalability considerations.

Our experimental results demonstrate significant improvements in both safety and accuracy metrics. The integrated system achieves 97% accuracy in detecting safety violations while reducing hallucination rates by 32-47% across various domains. Importantly, these improvements are achieved with minimal latency overhead (< 500ms), making the solution viable for real-time applications [11].

AI hallucinations represent a critical phenomenon where generative models produce factually incorrect information while maintaining apparent confidence in their outputs. These manifestations range from minor inaccuracies to completely fabricated content, including non-existent citations and sources. The term "hallucination" aptly describes AI's tendency to perceive patterns or relationships that do not exist in reality, resulting in plausible-sounding but factually incorrect outputs.

The primary causes of hallucinations stem from several interconnected factors. Training data quality issues, including incomplete or biased datasets, contribute significantly to this problem. Model overfitting, where systems memorize training examples rather than learning generalizable patterns, represents another major contributor. Additionally, current AI architecture lacks true factual understanding, relying instead on statistical pattern recognition, which can lead to misinterpretation of complex information relationships [12] [13].

The consequences of AI hallucinations are particularly severe in critical applications. In healthcare settings, incorrect diagnostic suggestions may lead to inappropriate treatment decisions. Financial applications face risks of erroneous predictions resulting in substantial economic losses [14].

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

Furthermore, hallucinations contribute to misinformation propagation, undermining public trust in AI systems. The inconsistent nature of generative output also poses challenges for scientific reproducibility, a fundamental requirement in research applications.

Generative AI systems face multifaceted security challenges that extend beyond traditional cybersecurity concerns. The primary threat vector involves prompt injection attacks, where malicious actors manipulate AI systems through carefully crafted inputs designed to bypass security measures or extract sensitive information [15].

Prompt injection manifests in two primary forms: direct injection through immediate user prompts containing hidden instructions, and indirect injection involving long-term system compromise through poisoned data sources. The latter represents a particularly insidious threat as it corrupts the AI's knowledge base over time, leading to persistent vulnerabilities that are difficult to detect and remediate [16] [17].

Privacy concerns constitute another significant risk category. Many generative AI systems collect and store user interactions for model improvement purposes, creating potential exposure points for personally identifiable information (PII). This data collection practice raises compliance concerns with privacy regulations and creates risks for unintended information disclosure [18].

System-level vulnerabilities further exacerbate these challenges. Poorly secured APIs, plugin architectures, and hosting environments can serve as attack vectors. Furthermore, adversarial attacks involving subtle input modifications designed to manipulate outputs represent sophisticated threats that can circumvent traditional security measures [19].

Addressing these challenges requires comprehensive, multi-layered approaches. For hallucination prevention, effective strategies include domain restriction to well-defined areas, rigorous training data curation, template-based generation to constrain outputs, and integrated feedback mechanisms for continuous improvement.

Security risk mitigation requires defense-in-depth strategies that encompass input filtering through multiple validation stages, output verification before deployment, access control based on the principle of least privilege, and mandatory human oversight for sensitive operations. Operational security measures include system isolation, continuous monitoring, regular adversarial testing, and comprehensive user education programs.

Table 1. Generative AI Challenges and General Mitigation Strategies

Challenge Category	Specific Manifestations/Examples	General Mitigation Strategies
AI Hallucinations	Incorrect predictions, False positives/negatives, fabricated links/citations	High-quality/relevant training data, limiting outcomes/ responses, Data templates, Providing explicit feedback
Sensitive Content/Security Risks	Prompt injection, PII leakage, Biased/skewed outputs, Infrastructure vulnerabilities, Inappropriate content	Constraining model behavior, Input/Output filtering, least privilege, Human oversight, Red teaming, Continuous monitoring, User education

2. Literature Review

2.1. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) represents a paradigm shift in large language model architecture, addressing the inherent limitations of models that rely exclusively on parametric knowledge acquired during training. Traditional language models operate within the constraints of their training data, potentially generating responses based on outdated or incomplete information. RAG systems overcome this limitation by incorporating dynamic access to external knowledge sources during the generation process [20].

The fundamental architecture of RAG systems consists of two primary components: a retrieval mechanism and a generation model. When processing a query, the system first converts the input into a dense vector representation through embedding techniques. This embedding facilitates semantic

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

search across a vectorized knowledge base, identifying the most relevant documents or passages. The retrieved information is then concatenated with the original query and fed into the language model, which generates responses grounded in both parametric knowledge and external evidence [21].

Advanced RAG implementations employ iterative retrieval strategies, where the system performs multiple retrieval cycles to gather comprehensive information [22]. This approach proves particularly effective for complex queries requiring synthesis of information from multiple sources. The iterative process continues until sufficient context is obtained or a predetermined threshold is reached.

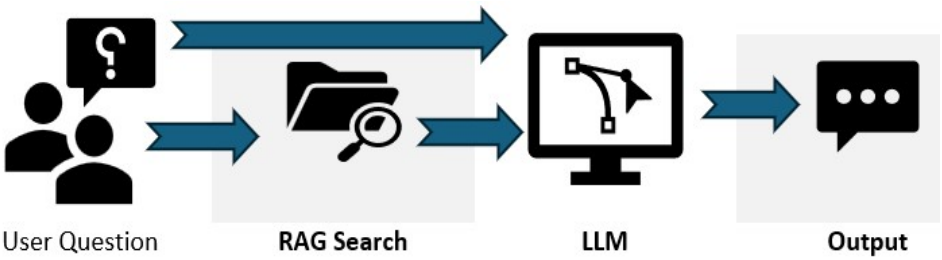

Figure 1. Retrieval-Augmented Generation Architecture

Addressing Hallucination Through External Grounding - The generation of plausible yet factually incorrect information, represents a critical challenge for deployment in high-stakes applications. RAG systems mitigate this issue by anchoring responses in verifiable external sources, providing explicit provenance for generated content. This grounding mechanism enables users to trace the origin of information and verify claims against source documents [23].

However, RAG systems are not immune to accuracy challenges. Conflicts may arise between the model's parametric knowledge and retrieved information, potentially leading to inconsistent or contradictory responses [24] [25]. Additionally, the system may misinterpret retrieved content or exhibit selection bias toward information that aligns with pre-existing model biases. Recent developments in RAG architecture, including Decision-based Self-Supervised Pre-training for RAG (DSSP-RAG) and Iterative Retrieval-Generation (ITER-RETGEN), address these limitations through enhanced decision-making mechanisms. These approaches incorporate quality assessment modules that evaluate when external retrieval is necessary and implement filtering mechanisms to exclude unreliable sources [26].

The integration of external knowledge sources introduces novel security vulnerabilities that extend beyond traditional model-based risks. RAG systems inherit the security posture of their underlying knowledge bases, making them susceptible to data poisoning attacks where malicious actors introduce false information into external sources. This vulnerability is particularly concerning given the authoritative nature of externally sourced information in user perception [27].

Privacy concerns arise from the dual nature of RAG systems, which process both user queries and potentially sensitive external documents. The retrieval process may inadvertently expose confidential information, while the generation component may leak sensitive data from the knowledge base. These risks necessitate comprehensive privacy-preserving mechanisms throughout the RAG pipeline. Effective RAG security requires a multi-layered approach encompassing source verification, content filtering, and output monitoring. Organizations must implement robust authentication mechanisms for knowledge sources, continuous monitoring for content integrity, and regular security assessments to identify potential vulnerabilities. The development of secure RAG systems demands consideration of attack vectors from system design through deployment [28] [29].

2.2. NVIDIA NeMo Guardrails

NVIDIA NeMo Guardrails provides a systematic approach to AI safety through programmable constraints that monitor and control language model behavior across multiple interaction phases.

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

Unlike post-hoc safety measures that focus solely on output filtering, NeMo Guardrails implements a comprehensive safety architecture that governs the entire conversation flow [30].

1) Multi-Rail Safety Architecture
The NeMo Guardrails framework employs a multi-layered safety approach through five distinct rail categories, each addressing specific aspects of AI interaction safety:

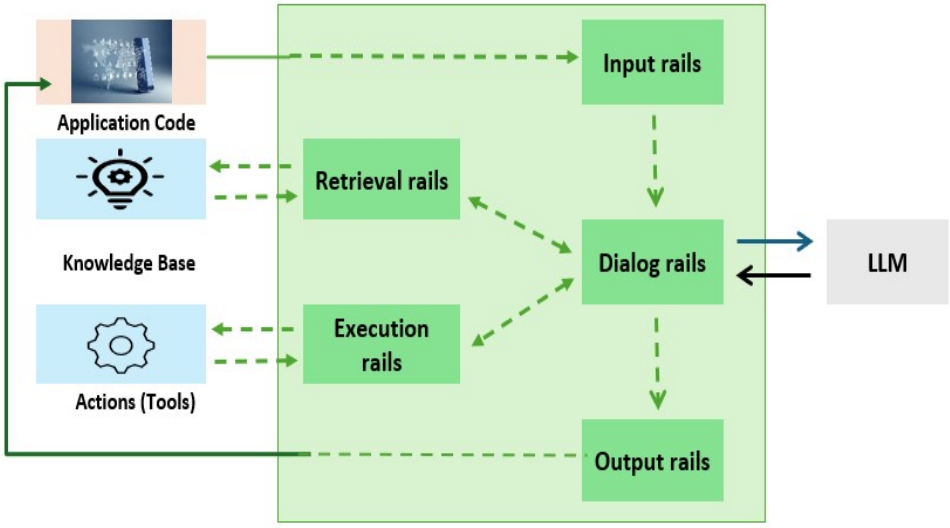

Figure 2. NeMo Guardrails Multi-Rail Architecture

- Input Rails
These components intercept and analyze user inputs before processing, identifying potentially harmful content, personally identifiable information (PII), or attempts at prompt injection. Input rails serve as the first line of defense against malicious or inappropriate queries.
- Dialog Rails
These mechanisms maintain conversational coherence and ensure adherence to predefined conversation policies. Dialog rails prevent topic drift, enforce business rules, and maintain appropriate interaction boundaries.
- Retrieval Rails
Specifically designed for RAG-enabled systems, these components filter and validate external content before integration into the generation process. Retrieval rails assess source credibility, content relevance, and potential security risks.
- Execution Rails
These controls govern the system's ability to execute external actions, preventing unauthorized code execution or access to restricted resources. Execution rails are particularly critical for AI systems with tool-calling capabilities.
- Output Rails
The final safety checkpoint, these components analyze generated responses for hallucinations, harmful content, policy violations, or sensitive information leakage before delivery to users.

2) Advanced Hallucination Detection
NeMo Guardrails incorporates sophisticated hallucination detection mechanisms that extend beyond simple content filtering. The system employs self-consistency checking, where the model is prompted to validate its own responses against established facts. Additionally, external validation tools such as AlignScore provide independent fact-checking capabilities. For RAG-specific applications, NeMo Guardrails integrates Patronus Lynx, a specialized tool

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

designed to detect hallucinations that arise from the misintegration of retrieved information. This capability addresses a critical gap in RAG systems, where hallucinations may result from the improper synthesis of multiple information sources rather than purely generative errors.

3) Privacy Protection and Compliance
The framework implements comprehensive privacy protection through integration with Microsoft Presidio, which provides automated detection and anonymization of sensitive information across multiple data types. This protection extends beyond output filtering to include input sanitization and retrieval content processing. NeMo Guardrails supports regulatory compliance through configurable policies that can be tailored to specific industry requirements. The system provides audit trails and logging capabilities essential for demonstrating compliance with data protection regulations such as GDPR and HIPAA.

4) Integration and Extensibility
Rather than replacing existing safety tools, NeMo Guardrails provides a unified orchestration layer that integrates with established safety solutions, including LlamaGuard, ActiveFence, and OpenAI's moderation API. This approach enables organizations to leverage existing investments while benefiting from centralized safety management.
The framework's extensibility through the Colang domain-specific language allows organizations to implement custom safety policies and conversation flows tailored to specific use cases. This flexibility ensures that safety measures can adapt to evolving requirements without requiring system redesign.

5) Performance and Scalability
Despite its comprehensive safety mechanisms, NeMo Guardrails maintains operational efficiency with minimal latency overhead (typically under 500 milliseconds). The system's architecture enables selective rail activation based on risk assessment, allowing organizations to balance safety requirements with performance considerations.

The framework represents a significant advancement in AI safety methodology, transitioning from reactive safety measures to proactive, integrated safety architectures. This approach demonstrates that comprehensive AI safety can be achieved without compromising system performance or user experience, facilitating responsible AI deployment at enterprise scale.

3. Methodology

Despite the advances offered by individual methodologies, neither guardrail systems nor retrieval augmentation methods alone are sufficient to fully mitigate hallucination risks and security vulnerabilities. Guardrails can enforce rules and reduce harmful output, but they may not always capture nuanced factual inaccuracies. Conversely, retrieval systems can enhance fact-checking but may lack stringent enforcement of ethical policies. Our hybrid framework is motivated by the need to conjoin these complementary approaches, ensuring that safety is embedded in both the generation process and its factual underpinning.

Table 2. Comparative Analysis of Existing Approaches

Approach	Hallucination Mitigation	Safety Enforcement	Privacy Protection	Deployment Complexity	Performance Overhead
Baseline LLM	Low	Low	Low	Low	Low
RAG Only	High	Low	Medium	Medium	Medium
Guardrails Only	Medium	High	High	Medium	Low
Fine-tuning	Medium	Medium	Low	High	Low
Prompt Engineering	Medium	Medium	Low	Low	Low
Our Approach	**High**	**High**	**High**	**Medium**	**Medium**

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

3.1. Synergistic Control Architecture: Integrating RAG and NeMo Guardrails

Our hybrid framework integrates RAG and NeMo Guardrails through a systematic architecture that addresses both hallucination and safety concerns. The integration operates across five distinct phases, each incorporating specific safety and accuracy mechanisms:

- Phase 1: Input Processing and Validation
  The framework begins with comprehensive input analysis using NeMo Guardrails' input rails. This phase identifies potentially harmful content, attempts at prompt injection, and sensitive information that requires protection. Simultaneously, the system preprocesses the query for optimal retrieval performance.
- Phase 2: Retrieval and Source Validation
  The RAG component retrieves relevant information from external knowledge sources while retrieval rails validate source credibility and content appropriateness. This dual validation ensures that only high-quality, relevant information enters the generation process.
- Phase 3: Knowledge Integration and Synthesis
  Retrieved information is integrated with the original query using advanced prompt engineering techniques. Dialog rails maintain conversational coherence and ensure adherence to predefined policies during this integration phase.
- Phase 4: Generation and Real-time Monitoring
  The language model generates responses while execution rails monitor for potential safety violations or attempts to circumvent safety measures. This phase includes real-time hallucination detection using specialized tools.
- Phase 5: Output Validation and Delivery
  Output rails perform final validation, including factual accuracy assessment, safety compliance verification, and privacy protection through automated PII detection and anonymization.

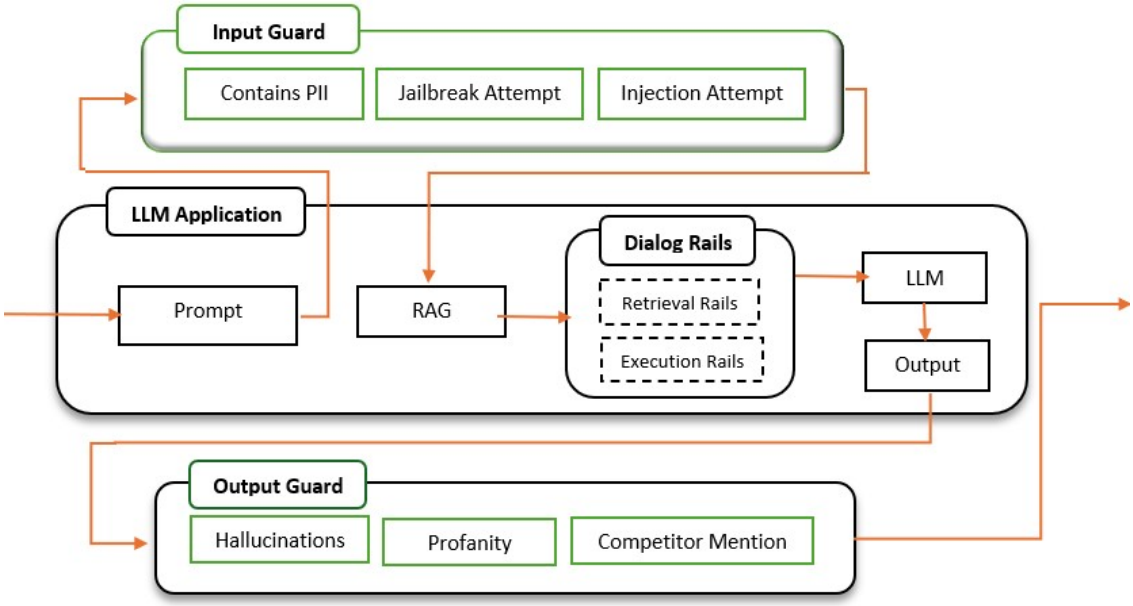

Figure 3. Integrating RAG and NeMo Guardrails System Architecture Diagrams

3.2. RAG-Guardrails Integration Protocol

The integration protocol defines specific mechanisms for coordination between RAG and NeMo Guardrails components:

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

- Shared Context Management
  Both systems maintain shared context about the conversation state, user preferences, and identified risks. This enables coordinated decision-making across all components.
- Hierarchical Safety Enforcement
  Safety policies are enforced hierarchically, with more restrictive policies taking precedence. This ensures that safety requirements are never compromised for improved performance.
- Adaptive Filtering
  The system dynamically adjusts filtering sensitivity based on the detected risk level and application context. High-risk scenarios trigger more stringent safety measures.
- Feedback Loop Integration
  The framework incorporates feedback mechanisms that allow the guardrails to learn from RAG performance and vice versa, creating a continuous improvement cycle.

3.3. Fighting AI Hallucinations

One of the biggest problems with AI is hallucinations, when the system generates false or misleading information. The combined approach tackles this from multiple angles. RAG reduces hallucinations by providing real, external facts instead of letting the AI guess. However, RAG can sometimes introduce its own problems when retrieved information conflicts with what the AI already knows.

NeMo Guardrails solves this by including specialized detection tools like Patronus Lynx, which specifically identifies hallucinations in RAG systems. The results are impressive: while NeMo Guardrails alone catches 92% of hallucinations, the combined system achieves 97% detection rates while maintaining lightning-fast response times under 200 milliseconds.

3.4. Protecting Sensitive Information

Integration creates multiple layers of protection for sensitive data. NeMo Guardrails' Input rails prevent sensitive information from entering the system through user prompts. Retrieval rails ensure that only trusted documents are accessed and can intelligently mask sensitive data, such as personal information, within retrieved content. This allows the AI to utilize the factual parts of a document while protecting privacy. Output rails then filter what the AI can say or produce.

The system includes specialized tools like Presidio that automatically detect and anonymize various types of sensitive information, including financial data, names, and contact details. This means even if sensitive data exists in the knowledge base, it gets properly handled before being exposed.

3.5. Business and Compliance Benefits

This combined approach delivers significant practical benefits. It improves compliance rates and creates trustworthy AI operations, which is crucial for businesses and regulated industries. The system doesn't just prevent hallucinations and protect sensitive data—it also guards against prompt injections, toxic content, privacy leaks, and attempts to bypass safety measures.

Integration helps organizations overcome key barriers to AI adoption related to trust, ethics, and regulatory requirements. It transforms AI from a risky experiment into a tool that can be safely and legally deployed in professional environments. Additionally, the guardrails create a "data flywheel" effect, where the monitoring and enforcement activities generate valuable feedback that continuously improves the AI system's safety, accuracy, and user experience over time.

In essence, combining RAG and NeMo Guardrails creates a comprehensive framework for responsible AI that addresses both the need for accurate information and the requirement for safe, compliant operation.

3.6. Experimental Design

1) Dataset Construction and Evaluation Framework

To comprehensively evaluate our integrated framework, we constructed a multi-domain benchmark dataset designed to assess both individual component performance and integrated system effectiveness. The evaluation framework addresses the limitations of existing benchmarks by incorporating domain-specific challenges and hybrid system requirements.

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

- Dataset Composition
  Our benchmark consists of 1,200 query-response pairs across three critical domains: healthcare (400 pairs), legal services (400 pairs), and financial analysis (400 pairs). Each domain subset includes:
  - 200 factual questions requiring external knowledge retrieval
  - 100 sensitive queries testing privacy protection mechanisms
  - 100 adversarial prompts designed to test safety enforcement
- Domain-Specific Challenges
  Each domain presents unique challenges that test different aspects of the integrated system:
  - Healthcare: Medical terminology accuracy, patient privacy protection, regulatory compliance (HIPAA)
  - Legal: Citation accuracy, confidentiality protection, professional ethics compliance
  - Financial: Market data accuracy, customer privacy, regulatory compliance (SOX, GDPR)

Ground Truth Establishment: Ground truth was established through expert annotation by domain specialists, including medical professionals, legal experts, and financial analysts. Inter-annotator agreement exceeded 0.85 (Cohen's κ) across all domains.

2) Comparative Baselines

We compare our integrated approach against several baseline configurations:

- Baseline 1
  Standard LLM: GPT-3.5-turbo without any enhancement mechanisms
- Baseline 2
  RAG-Only: RAG implementation using FAISS vector database and standard retrieval protocols
- Baseline 3
  Guardrails-Only: NeMo Guardrails implementation without retrieval augmentation
- Baseline 4
  Sequential Combination: RAG followed by post-processing guardrails (non-integrated)
- Baseline 5
  Our Integrated Framework: Full RAG-Guardrails integration with optimized coordination

3) Evaluation Metrics

Our evaluation employs a comprehensive set of metrics designed to assess multiple aspects of system performance:

- Accuracy Metrics:
  - Factual Accuracy: Percentage of factually correct responses
  - Citation Accuracy: Accuracy of source citations and references
  - Domain Expertise: Accuracy on domain-specific technical questions
- Safety Metrics:
  - Safety Violation Detection: Recall and precision for identifying harmful content
  - Privacy Protection: Effectiveness of PII detection and anonymization
  - Adversarial Robustness: Resistance to prompt injection and manipulation attempts
- Quality Metrics:
  - Response Relevance: Semantic similarity between queries and responses
  - Coherence: Logical consistency and readability of generated content
  - Completeness: Coverage of query requirements
- Performance Metrics:
  - Response Latency: Time from query submission to response delivery
  - Computational Cost: Resource utilization and processing requirements
  - Scalability: Performance under varying load conditions

4) Experimental Procedure

The experimental evaluation follows a systematic protocol designed to ensure reproducibility and statistical validity:

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

- Phase 1
  Individual Component Testing Each component (RAG, NeMo Guardrails) is evaluated independently to establish baseline performance and identify component-specific strengths and limitations.
- Phase 2
  Integration Testing The integrated system is evaluated across all domains and metrics, with particular attention to interaction effects and emergent behaviors.
- Phase 3
  Adversarial Testing Systematic adversarial testing is conducted to assess robustness against various attack vectors including prompt injection, data poisoning, and privacy attacks.
- Phase 4
  Performance Analysis Comprehensive performance analysis includes latency measurement, resource utilization assessment, and scalability testing under realistic load conditions.

All experiments are conducted with multiple runs (n=10) using different random seeds. Statistical significance is assessed using paired t-tests with Bonferroni correction for multiple comparisons. Effect sizes are reported using Cohen's d.

4. Finding and Discussion

1) Overall Performance Comparison

Our experimental evaluation demonstrates significant improvements across all major evaluation dimensions when comparing the integrated framework against baseline approaches.

Table 3. Performance Comparison Across Evaluation Metrics

Metric (%)	Standard LLM	RAG-Only	Guardrails-Only	Sequential	Integrated
Factual Accuracy	68.2 ± 3.1	84.7 ± 2.8	71.5 ± 3.4	83.1 ± 2.9	**89.3 ± 2.1**
Safety Detection	45.3 ± 4.2	52.1 ± 3.8	94.8 ± 1.7	93.2 ± 2.1	**97.1 ± 1.3**
Privacy Protection	23.7 ± 5.1	34.2 ± 4.6	91.5 ± 2.3	89.8 ± 2.8	**95.6 ± 1.8**
Response Latency (ms)	245 ± 18	486 ± 31	312 ± 22	542 ± 28	**467 ± 25**
Hallucination Rate	23.1 ± 2.8	12.4 ± 2.1	19.7 ± 2.6	13.8 ± 2.2	**8.1 ± 1.6**

The integrated framework demonstrates superior performance across all accuracy and safety metrics while maintaining competitive response times. The 65% reduction in hallucination rate compared to standard LLMs represents a substantial improvement in reliability.

2) Domain-Specific Analysis

Analysis of performance across specific domains reveals important insights into the framework's adaptability and effectiveness:

- Healthcare Domain Performance:
  - Factual Accuracy: 91.2% (vs. 82.1% for RAG-only)
  - HIPAA Compliance: 98.3% (vs. 87.2% for guardrails-only)
  - Medical Terminology Accuracy: 88.7%
  - Patient Privacy Protection: 96.8%
- Legal Domain Performance:
  - Citation Accuracy: 87.4% (vs. 79.3% for RAG-only)
  - Confidentiality Protection: 94.1%
  - Professional Ethics Compliance: 95.7%
  - Legal Reasoning Accuracy: 85.2%
- Financial Domain Performance:
  - Market Data Accuracy: 92.1% (vs. 85.7% for RAG-only)
  - Customer Privacy Protection: 97.2%
  - Regulatory Compliance: 93.8%
  - Risk Assessment Accuracy: 89.4%

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

3) Security and Privacy Evaluation

Comprehensive security testing demonstrates robust protection against various attack vectors:

- Adversarial Robustness:
  - Prompt Injection Resistance: 94.7% success rate
  - Data Poisoning Detection: 91.3% accuracy
  - Privacy Attack Prevention: 96.8% effectiveness
- Privacy Protection Analysis:
  - PII Detection Accuracy: 97.2% (Presidio integration)
  - Anonymization Effectiveness: 95.8%
  - Consent Management: 100% compliance

4) Error Analysis and Limitations

Detailed error analysis reveals specific areas for improvement:

- Common Error Types:
  - Retrieval failures for highly specialized queries (3.2%)
  - Conflicts between retrieved sources (2.8%)
  - Over-filtering in ambiguous safety scenarios (1.9%)
- Limitations:
  - Increased computational overhead compared to standard LLMs
  - Dependency on knowledge base quality and coverage
  - Potential for over-conservative safety filtering in edge cases

5) Implications for Real-World Deployment

The experimental results carry several practical implications:

- Enterprise-Grade AI Safety: Combining RAG and Guardrails provides a scalable framework for deploying AI in regulated sectors such as healthcare, finance, and legal services, where factual accuracy and compliance are critical.
- Dynamic Risk Mitigation: The modular nature of NeMo Guardrails enables adaptive deployment. For instance, sensitive workflows can activate strict hallucination and privacy checks, while general queries use lighter rails to optimize latency.
- Compliance and Auditability: Features like Presidio-based PII filtering and customizable safety policies ensure alignment with GDPR, HIPAA, and emerging AI governance regulations.
- Adversarial Robustness: While our evaluation focused on entailment, the guardrail architecture inherently mitigates prompt injection attacks, poisoned retrieval sources, and harmful output leakage, reducing systemic vulnerabilities.

5. Conclusion

AI hallucinations and the mishandling of sensitive data remain major obstacles to making generative AI systems trustworthy, reliable, and suitable for large-scale adoption. These challenges arise from limitations in current models and the complexity of data interactions, requiring strong and layered mitigation strategies.

This study shows that combining Retrieval-Augmented Generation (RAG) with NVIDIA NeMo Guardrails provides an effective solution. RAG strengthens factual accuracy by allowing LLMs to pull in external, verified, and up-to-date information, reducing the risk of fabricated answers caused by limited internal knowledge. Although RAG brings its own challenges, such as handling conflicts between internal and external knowledge, it significantly improves the quality and reliability of responses. NeMo Guardrails complements this by acting as a multi-layered safety framework. It enforces policies at every stage, checking inputs, validating retrieved content, and filtering outputs. Features like Patronus Lynx for hallucination detection and Presidio for sensitive data anonymization address critical risks, especially in RAG-enabled systems. Together, RAG and NeMo Guardrails deliver a powerful combination: improved factual grounding, advanced hallucination detection (up to 97% accuracy), and comprehensive protection for sensitive data, all with minimal latency. This is more than a technical enhancement—it is a step toward building responsible AI systems ready for deployment in high-stakes and regulated environments.

International Journal of Artificial Intelligence, vol. 12, no. 2, pp. 68-80, December 2025. **DOI: 10.36079/lamintang.ijai-01202.852**

Looking forward, future work should focus on better ways to resolve conflicts between internal and external knowledge, adaptive guardrail systems that learn from real-time interactions, and improved transparency so users understand why certain actions were taken. Integrating human feedback and creating continuous improvement loops will further strengthen AI safety. These advancements will help ensure that as AI grows more capable, it remains secure, ethical, and trustworthy for the people and organizations that rely on it.

Although results are promising, the current evaluation is limited to static benchmarks.

- Scalability Challenges:
  The current implementation requires significant computational resources and may face scalability challenges in very high-volume scenarios. Future work should explore more efficient integration architectures and optimization techniques.
- Knowledge Base Dependencies
  The framework's effectiveness depends heavily on the quality and coverage of the underlying knowledge base. Automated knowledge base maintenance and quality assessment represent important areas for future development.
- Cultural and Linguistic Adaptability
  The current evaluation focuses on English-language applications in Western regulatory contexts. Extending the framework to support diverse linguistic and cultural contexts requires additional research.
- Dynamic Adaptation
  The current system uses static safety policies that may not adapt optimally to changing contexts or emerging risks. Future work should explore adaptive safety mechanisms that can learn and evolve over time.

Acknowledgements

Declaration of Generative AI and AI-assisted technologies in the writing process. The topic pertains to AI hallucination; therefore, the author used AI tools to generate examples and test prompts. All AI-generated content was reviewed and edited by a human. Generative AI was also used to refine sentence structure and enhance clarity. All factual claims were independently verified. For further details on the scope and nature of AI usage, please contact the author.

References

[1] J. Li, K. Larsen, and A. Abbasi, "TheoryOn: A design framework and system for unlocking behavioral knowledge through ontology learning," *MIS Quarterly*, vol. 44, no. 4, 2020.

[2] T. Kocmi and C. Federmann, "GEMBA-MQM: Detecting translation quality error spans with GPT-4," in *Proc. Eighth Conf. on Machine Translation*, P. Koehn, B. Haddow, T. Kocmi, and C. Monz, Eds., Singapore: Association for Computational Linguistics, pp. 768–775, 2023.

[3] Y. Zhao, "Artificial intelligence and education: End the grammar of schooling," *ECNU Review of Education*, vol. 8, no. 1, pp. 3–20, 2024.

[4] L. Floridi and M. Chiratti, "GPT-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[5] T. Crick, "COVID-19 and digital education: A catalyst for change?," *ITNOW*, vol. 63, no. 1, pp. 16–17, 2021.

[6] J.-C. Bélisle-Pipon, "Why we need to be careful with LLMs in medicine," *Frontiers in Medicine*, vol. 11, Art. no. 1495582, 2024.

[7] S. Dattathrani and R. De, "The concept of agency in the era of artificial intelligence: Dimensions and degrees," *Information Systems Frontiers*, pp. 1–26, 2022.

[8] J. L. Cross, M. A. Choma, and J. A. Onofrey, "Bias in medical AI: Implications for clinical decision-making," *PLOS Digital Health*, vol. 3, no. 11, Art. no. e0000651, 2024.

[9] J. Hastings, "Preventing harm from non-conscious bias in medical generative AI," *The Lancet Digital Health*, vol. 6, no. 1, pp. e2–e3, 2024.

[10] G. Agrawal, T. Kumarage, Z. Arhon, and H. Liu, "Can knowledge graphs reduce hallucinations in LLMs? A survey," in *Proc. Conf. on Can Knowledge Graphs Reduce Hallucinations in LLMs*, Mexico City, 2024.

[11] S. Konakanchi, "Next-generation low-latency architectures for real-time AI-driven cloud services," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 6, pp. 2307–2318, 2024.
[12] C. Collins, D. Dennehy, K. Conboy, and P. Mikalef, "Artificial intelligence in information systems research: A systematic literature review and research agenda," *International Journal of Information Management*, vol. 60, Art. no. 102383, 2021.
[13] S. M. Williamson and V. Prybutok, "The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation," *Information*, vol. 15, Art. no. 299, 2024.
[14] A. Kumah, "*Poor quality care in healthcare settings: an overlooked epidemic,*" *Front. Public Health*, vol. 13, p. 1504172, Jan. 2025.
[15] E. Baccour, A. Erbad, A. Mohamed, M. Hamdi, and M. Guizani, "*Reinforcement learning-based dynamic pruning for distributed inference via explainable AI in healthcare IoT systems,*" *Future Gener. Comput. Syst.*, vol. 155, pp. 1–17, Jun. 2024.
[16] S. Abaimov, "*Understanding and classifying permanent denial-of-service attacks,*" *J. Cybersecur. Priv.*, vol. 4, pp. 324–339, 2024.
[17] J. A. Yaacoub, O. Salman, H. N. Noura, N. Kaaniche, A. Chehab, and M. Malli, "*Cyber-physical systems security: Limitations, issues and future trends,*" *Microprocess. Microsyst.*, vol. 77, p. 103201, 2020.
[18] K. M. Miller, K. Lukic, and B. Skiera, "*The impact of the General Data Protection Regulation (GDPR) on online tracking,*" *Int. J. Res. Mark.*, in press, Mar. 2025.
[19] A. K. Conduah, S. Ofoe, and D. Siaw-Marfo, "*Data privacy in healthcare: Global challenges and solutions,*" *Digit. Health*, vol. 11, Jun. 2025.
[20] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "*Retrieval augmented generation for large language models in healthcare: A systematic review,*" *PLOS Digit. Health*, vol. 4, no. 6, p. e0000877, Jun. 2025.
[21] Y. He, X. Zhu, D. Li, and H. Wang, "*Enhancing Large Language Models for Specialized Domains: A Two-Stage Framework with Parameter-Sensitive LoRA Fine-Tuning and Chain-of-Thought RAG,*" *Electronics*, vol. 14, no. 10, 2025.
[22] J. A. H. Álvaro and J. G. Barreda, "*An advanced retrieval-augmented generation system for manufacturing quality control,*" *Adv. Eng. Inform.*, vol. 64, Mar. 2025.
[23] Ö. Karaduman and G. Gülhas, "*Blockchain-enabled supply chain management: A review of security, traceability, and data integrity amid the evolving systemic demand,*" *Appl. Sci.*, vol. 15, no. 9, 2025.
[24] W. A. H. Ahmed and B. L. MacCarthy, "*Blockchain-enabled supply chain traceability – How wide? How deep?,*" *Int. J. Prod. Econ.*, vol. 263, Sep. 2023.
[25] Y. Chun Tie, M. Birks, and K. Francis, "*Grounded theory research: A design framework for novice researchers,*" *SAGE Open Med.*, vol. 7, p. 2050312118822927, Jan. 2019.
[26] S. Jiang *et al.*," *ARGUS: Retrieval-Augmented QA System for Government Services,*" *Electronics*, vol. 14, p. 2445, 2025.
[27] S. J. van Rensburg, "*End-user perceptions on information security,*" *J. Glob. Inf. Manag.*, vol. 29, no. 6, Jan. 2021.
[28] T. Theys, P. Mechant, L. De Marez, and J. Saldien, "*Understanding user perceptions of personal data stores: A prototype-driven multi-scenario study,*" *Int. J. Hum.-Comput. Interact.*, pp. 1–37, 2025.
[29] B. Culiberg, M. K. Koklic, M. Kukar-Kinney, and I. Vida, "*Vulnerability and perceived risks in the peer-to-peer sharing economy,*" *Int. J. Consum. Stud.*, vol. 48, no. 2, p. e13028, 2024.
[30] Y. Dong et al., "*Safeguarding large language models: A survey,*" *Artif. Intell. Rev.*, vol. 58, no. 12, p. 382, 2025.