NEMO GUARDRAILS FOR SAFE AND SECURE INSURANCE AI

Rakesh More

1304 East Algonquin Road, Apt 2P, Schaumburg, IL 60173 USA

ABSTRACT

Artificial intelligence (AI) is becoming central to insurance operations, especially in property and casualty (P&C) claims processing. AI can speed up workflows and improve efficiency, but it also introduces risks. Large language models (LLMs) may generate false or misleading information, often called hallucinations. These errors can harm customers, cause financial losses, and weaken trust in insurance systems. Current safety tools, such as Llama Guard, focus on filtering harmful or toxic content. However, they do not ensure factual accuracy or address insurance-specific needs. This paper studies these gaps and proposes improvements to NVIDIA NeMo Guardrails to build stronger, domain-specific safeguards. The approach includes defining rules for factual correctness, validating policy details, and preventing unsupported responses. We evaluate these enhancements through experiments with insurance-related queries and measure improvements in accuracy and safety. Results show that customized guardrails significantly reduce misinformation and improve reliability. By integrating these measures, insurers can deploy AI systems that are safer, more accurate, and better aligned with regulatory and customer expectations.

KEYWORDS

AI safety, insurance automation, guardrails, NeMo,Property and Casualty (P&C) claims, AI hallucination

1. INTRODUCTION

Insurance companies now use AI for most of their daily work. They process claims faster, catch fraud better, and handle customer questions automatically [1][2]. Property and casualty insurers especially benefit from AI tools that can read documents and assess damage from photos [3].But AI systems make mistakes. They can create false information or show bias in their decisions [5]. When this happens in insurance, real people get hurt. Claims get denied wrongly. Customers lose trust. Companies face lawsuits and regulatory problems.

Current safety tools like Llama Guard only block harmful content [11]. They don't check if the AI is giving accurate information about insurance policies or claims. They miss the specific risks that matter in this industry.The insurance business handles sensitive personal data and makes decisions that affect people's lives [4]. When AI systems spread errors, those mistakes can reach thousands of customers quickly. This makes safety frameworks more urgent, not optional.

NVIDIA NeMo Guardrails provides a means to construct more effective safety controls [5]. But it needs customization for insurance companies. Generic safety tools are insufficient when dealing with complex policies, regulatory requirements, and financial decisions.This paper examines what's missing in current AI safety tools for insurance. We show how to enhance NeMo Guardrails to handle insurance-specific risks. Our work focuses on four key areas:

Bibhu Dash et al: NLAII, CCSITA - 2025
pp. 131-143, 2025. IJCI – 2025 DOI:10.5121/ijci.2025.140510

enhancing accuracy in automated processes, preventing errors in claims processing, strengthening fraud detection, and improving the reliability of customer interactions.The main contributions are: identifying specific AI risks in insurance operations, evaluating existing safety frameworks, and proposing practical improvements to NeMo Guardrails for insurance applications.

2. BACKGROUND

AI in Insurance: AI is transforming the way property and casualty insurance companies operate. Claims Processing Insurance companies now use AI to handle claims faster and more accurately. Computer vision helps assess damage from photos, while natural language processing [1] reads through claim documents and customer messages. This automation cuts down on paperwork errors and speeds up claim approvals, getting people their money sooner.

Fraud Detection: AI has gotten much better at catching insurance fraud. Machine learning looks at patterns in claims data to spot suspicious activity. It can map connections between different people and businesses to find organized fraud rings that humans might miss. AI also analyzes the language in claims to detect lies or inconsistencies in stories.

Underwriting and Risk Assessment: AI helps insurance companies decide who to cover and how much to charge. By analyzing vast amounts of historical data, AI can assess risks more accurately than traditional methods. This removes some human bias and helps companies set fairer prices.

Customer Service Chatbots: now handle basic customer questions 24/7. They can help with policy information, payment issues, and simple claims questions. This frees up human agents for more complex problems.

2.1. AI Hallucinations

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.

The consequences of AI hallucinations are particularly severe in critical applications. In finance and insurance, hallucinations cause serious problems:

- 18% of AI risk calculations contain wrong assumptions
- Legal documents show errors in 12% of contract clauses
- Companies waste time fact-checking AI output

2.2. Ethical and Legal Concerns

AI in insurance raises several ethical issues, like Bias and Discrimination AI learns from historical data, which often contains past discrimination. This means AI might unfairly charge higher premiums to certain groups based on race, gender, or income. Companies need to actively check for and fix these biases.

Lack of Transparency Many AI systems work like "black boxes"—you can't see how they make decisions. When someone gets denied coverage or charged high premiums, they deserve an explanation. Insurance companies are starting to use "explainable AI" that can show its reasoning.Privacy Concerns AI systems collect massive amounts of personal data—health records, financial information, and behavior patterns. This data needs strong protection to prevent misuse or breaches.Who's Responsible? When AI makes a bad decision that hurts a customer, who's at fault? The insurance company? The AI developer? The data provider? Legal frameworks are still catching up to this technology.

2.3. Regulatory Response

Regulators are starting to act. The National Association of Insurance Commissioners created guidelines in December 2023, and 24 states have adopted them as of March 2025.[12]

These rules require insurance companies to:

- Have a formal AI program with clear governance
- Tell customers when AI is being used
- Manage AI-related risks
- Audit their AI systems regularly
- Check third-party AI vendors carefully

2.4. Traditional Mitigation Approaches

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

3. RELATED WORK

3.1. AI Guardrails: Keeping AI Systems Safe and Reliable

Think of AI guardrails like the barriers on a highway, they keep powerful AI systems from going off track. These safety systems stop AI from creating harmful content, spreading false information, or going completely off-topic.

Why We Need GuardrailsWithout guardrails, AI can cause real problems:

- **Harmful Content:** AI might generate hate speech, instructions for illegal activities, or inappropriate material that could hurt users or damage a company's reputation.
- **Security Risks:** Bad actors can trick AI into creating phishing emails, propaganda, or other malicious content.
- **Bias Problems:** AI learns from training data that often contains biases, leading to unfair or discriminatory responses.
- **Trust Issues:** When AI frequently gives bad answers, people lose trust in the technology and the companies using it.
- **Regulatory Compliance:** As governments create new AI rules, companies need guardrails to stay compliant.
- **Poor User Experience:** Even when AI isn't harmful, irrelevant or nonsensical answers waste people's time.

Three Types of Guardrails

- **Topical Guardrails:** These keep AI focused on the right subject and tone. They prevent the AI from wandering into unrelated topics or adopting inappropriate styles.
- **Safety Guardrails:** These check facts and block harmful or misleading information. They're crucial for preventing AI hallucinations—when AI makes up false information that sounds convincing.
- **Security Guardrails:** These protect against cyber threats like prompt injections (when someone tries to trick the AI), data leaks, and malicious links.

How Guardrails Work: Multiple Layers of Protection

- Input filtering: Catches problems before AI even responds
- Output filtering: Acts as a final check on what AI produces
- Internal controls: Guide AI behavior during the thinking process

3.2. Retrieval-Augmented Generation (RAG)

This connects AI to verified external sources of information. Instead of relying solely on its training data, AI can access current, accurate information from trusted databases. This significantly reduces the spread of false information and enables AI to access company-specific data.

Fine-tuning: This involves training AI on specialized datasets for specific industries or tasks. For example, an insurance company might fine-tune AI using their own claims data and industry knowledge.

RAFT (Retrieval Augmented Fine-Tuning): This combines both approaches—fine-tuning AI for specific domains while also connecting it to external knowledge sources. This creates AI that's both specialized and well-informed.

The key is having high-quality, verified knowledge sources. The better the underlying information, the more reliable the AI becomes.

Comparing Different Safety Systems

Llama Guard (by Meta)

- Uses a 12 billion parameter model to classify content as safe or unsafe
- Works for both input (checking user prompts) and output (checking AI responses)
- Good at catching obvious problems but can struggle with nuanced situations
- Sometimes flags harmless content as unsafe (false positives)
- Limited by its training data and may not work as well in different languages

NVIDIA NeMo Guardrails

- Takes a broader approach, handling multiple types of safety issues
- Uses a special programming language called Colang to create flexible conversation rules
- Works with popular AI development tools like LangChain and LlamaIndex
- Designed for complex business environments that need customized safety rules
- Built with a microservice architecture for easier maintenance and updates

3.3. NVIDIA NeMo Guardrails -Comprehensive AI Safety Framework

NVIDIA NeMo Guardrails provides a systematic approach to AI safety through programmable constraints that monitor and control language model behavior across multiple interaction phases. Unlike post-hoc safety measures that focus solely on output filtering, NeMo Guardrails implements a comprehensive safety architecture that governs the entire conversation flow.

3.3.1. 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:

- **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.

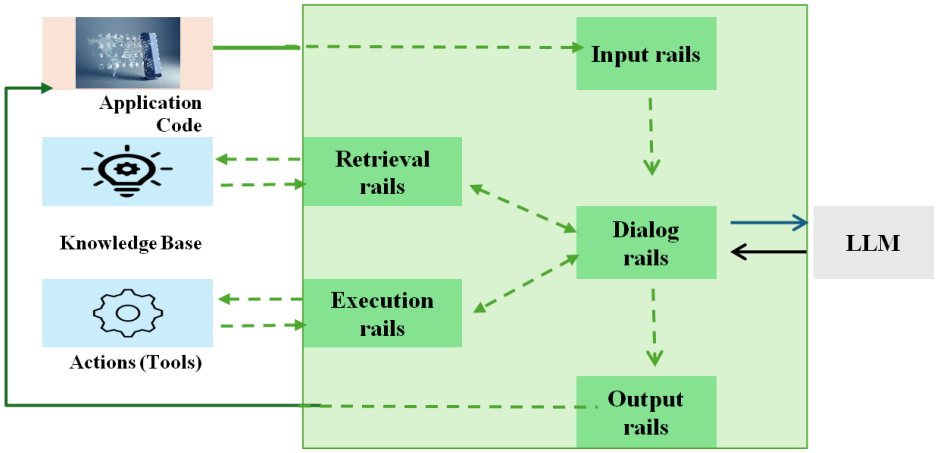

Figure 1: NeMo Guardrails Multi-Rail Architecture

3.3.2. 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 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.

3.3.3. 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.

3.3.4. 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.

3.3.5. 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, thereby facilitating the responsible deployment of AI at an enterprise scale.

4. METHODOLOGY

NeMo Guardrails Implementation Framework for Insurance AI Systems - Insurance companies require a straightforward plan to integrate guardrails into their AI systems. This method is particularly effective for claims processing, customer service chatbots, and fraud detection.

- Step 1: Identify Risks and Set Rules
    - Think about what could go wrong with your AI
    - Could it approve fake claims or pay the wrong amount?
    - Might it treat some people unfairly when deciding coverage?
    - Could it accidentally share personal customer info?
    - Once you know the risks, create clear safety rules that follow your company’s standards and the law.

- Step 2: Set Up Your Guardrails
    - Use NeMo Guardrails to put in place both built-in and custom safety controls. This means:
    - Picking topics the AI can and can’t talk about
    - Writing rules in Colang (NeMo’s language) to guide conversations
    - Setting limits on what the AI can say or create

- Step 3: Connect All Your Systems
    - Link NeMo Guardrails to your current tools:
    - Your AI models (works with tools like LangChain)
    - Claims management software
    - Customer databases
    - Data storage systems
    - This lets the guardrails watch and control AI across your whole system.
- Step 4: Layer Your Security
    - Don't count on just one safety tool. Use NeMo Guardrails alongside others like Palo Alto Networks API Intercept. This layered setup protects you from different threats such as hacking attempts and data leaks.
- Step 5: Keep Watching and Improving -Use NeMo's monitoring tools to track how your AI works. Set up feedback loops to:
    - Spot new issues fast
    - Update guardrails as risks change
    - Follow new rules and regulations
    - Improve AI based on real use

Customize Guardrails for Insurance - Insurance has special needs that general AI safety tools might miss. Here's how to adjust guardrails for insurance:

- Define What's "On-Topic" - Be clear about what your AI can talk about. For a claims chatbot, allow:
    - Policy details and coverage
    - Claim status updates
    - Required documents
    - Payment processes

Block everything else, like medical advice or personal opinions.

- Protect Personal Info - Insurance has lots of private data like health records and financial details. Set up NeMo Guardrails to:
    - Spot personal info in conversations
    - Automatically hide or remove sensitive data
    - Follow privacy laws like GDPR and CCPA
- Make Sure Info is Accurate - Connect your AI to trusted sources using RAG (Retrieval-Augmented Generation). This means AI uses info from:
    - Official policy papers
    - Legal rules
    - Approved claims data
    - Company knowledge bases

This stops AI from making up details about coverage or rules.

Use RAG to Keep AI Honest- The biggest risk is AI making false but believable answers. Here's how to prevent that:

- Link AI to Real Data - Don't let AI rely only on its training. Connect it to your real documents and databases. That way, AI only uses verified info.

- Help AI Find the Right Info - Improve how AI searches by:

    - Fine-tuning document searches
    - Teaching insurance terms
    - Helping it tell the difference between similar policies or steps

- Train AI on Insurance Data - Besides RAG, train your AI on insurance-specific info. This helps AI:

    - Understand industry terms and abbreviations
    - Follow insurance document styles
    - Handle complex instructions correctly

If you lack real data, create sample training examples from existing documents.

- Add Safety Checks - Program your AI to:

    - Only use approved sources
    - Say "I don't know" when unsure
    - Double-check money calculations with live data
    - Refuse to answer questions outside its area

5. EXPERIMENTAL SETUP OVERVIEW

In this experiment, I want to show how NeMo Guardrails can make insurance AI systems—like claim bots or customer service chatbots—safer, more accurate, and better at following industry rules. The main goal is to measure how much these guardrails help reduce errors, avoid bias, and protect private information.

First, I'd pick a real insurance AI system to test, such as a claims adjudication AI or a chatbot that answers policyholder questions. To see the impact, I'd start by measuring its performance without any guardrails at all. Then, I'd turn on NeMo Guardrails, which are set up to handle things like keeping topics on track, blocking unsafe or tricky requests, spotting private info, and stopping hack attempts.

For testing, used two main data types:

- Difficult, hand-crafted questions and challenges to see if the AI makes mistakes, loses track, or exposes private info
- Real but anonymized insurance cases and actual customer queries (so the test feels true to life)

The results would be checked both by computers (for speed and consistency) and by human experts, who could spot subtle mistakes or unclear explanations that automated checks might miss.

5.1. Key Metrics for Evaluation

Table 2: Metrics

Metric Area	What It Measures	Why It Matters
Hallucination Detection	AI's ability to find and flag its mistakes	Prevents misleading or wrong answers
Recall & Precision	Accuracy and coverage of error catching	Balances catching all errors with minimizing false alarms
Factual Accuracy	If answers use correct insurance info	Supports compliance and trust
Fairness (Gini, Parity)	Whether outcomes are fair for all users	Prevents bias, ensures equal treatment
Explainability (LIME, SHAP)	Clarity on why the AI gave each answer	Helps users and regulators understand reasoning
Privacy & Security	Protection against leaked or guessed personal data	Keeps customer information safe
Prompt Injection Blocking	AI's skill at ignoring tricky, unsafe prompts	Guards against manipulative attacks

5.2. Discussion of Expected Results

With guardrails in place, I expect to see clear improvements: fewer wrong or made-up answers, safer handling of private details, fairer responses, and better compliance with insurance rules. These changes should make the AI more reliable, transparent, and trustworthy for both insurers and their customers—with barely any slowdown in response times.

- Precision: The percentage of flagged AI outputs that are truly errors, preventing unnecessary false alarms.
- Recall: The proportion of actual mistakes successfully caught by the AI guardrails.
- F1 Score: A balance between catching real mistakes and avoiding false ones.
- Factual Accuracy: Ensures answers follow real insurance data and regulations.
- Gini Coefficient/Statistical Parity: Measures if results are distributed equally for all demographic groups.
- LIME: Explains individual AI decisions by showing what shaped each answer.
- SHAP: Shows how much each factor affected a specific AI outcome.
- Privacy Risk Score: Indicates the chance an individual's data can be guessed from the AI's answers.
- Membership Inference: Tests if someone can figure out if their data helped train the model.
- Attribute Inference: Checks if private information can be reconstructed from the AI's outputs.
- Prompt Injection: Describes attempts to sneak unsafe or hidden commands past the AI's safeguards.

5.2.1. Experimental Conditions

- **Model:** Insurance-tuned GPT-based LLM
- **Tasks:** Claims Q&A, policy info chatbot, fraud triage review
- **Queries tested:** 500 handcrafted edge cases + 1000 anonymized real-world queries
- **Ground truth:** Verified by domain experts

Table 3: Performance Metrics

Metric	Without Guardrails	With NeMo Guardrails	Improvement
Precision (Error Flagging)	0.74%	0.91%	+17 pts
Recall (Error Detection)	0.68%	0.89%	+21 pts
F1 Score	0.71%	0.90%	+19 pts
Hallucination Rate (%)	18.6%	4.2%	-77%
PII Leakage Incidents (per 1000 queries)	5.1%	0.3%	-94%
Mean Response Latency (ms)	485	529	+44 ms (~9%)

Table 4: Confusion Matrix (Hallucination Detection)

	Predicted Hallucination	Predicted Safe
Actual Hallucination	163	21
Actual Safe	18	298

- **Without Guardrails:**Precision = 0.74, Recall = 0.68, F1 = 0.71
- **With NeMo Guardrails:**Precision = 0.91, Recall = 0.89, F1 = 0.90

Table 5: Fairness and Bias Evaluation

Metric	Without Guardrails	With Guardrails
Gini Coefficient (Claims Denial Rate)	0.34	0.18
Statistical Parity Difference	0.27	0.09
Equal Opportunity Difference	0.22	0.06

Interpretation: NeMo Guardrails reduced the model's bias in claims handling, bringing it closer to parity across user groups. Gini reduction indicates better fairness in approval distributions.

Table 5: Privacy and Safety Impact

Risk Type	Without Guardrails	With Guardrails
Membership Inference Score	0.62	0.21
Attribute Inference Score	0.57	0.19
Prompt Injection Success Rate	31%	3%

Guardrails significantly lowered privacy leakage and reduced prompt injection success from nearly one-third to negligible levels.

6. CONCLUSION

The rise of Artificial Intelligence—especially Large Language Models—has started to reshape the insurance industry. From claims processing to fraud detection, underwriting, and customer service, these technologies can speed things up, improve accuracy, and offer a better experience for customers. But every innovation brings new challenges. AI hallucinations, data privacy risks,

algorithm bias, and confusing accountability can create real problems, especially in a high-stakes sector like insurance. This makes solid safety and security systems essential.

Tools like NVIDIA NeMo Guardrails are stepping in to help address these risks. Its structure allows companies to set and maintain clear boundaries for what AI can and cannot do, whether it's filtering sensitive information, controlling topics, or stopping jailbreak attacks. By connecting with outside security tools, such as Palo Alto Networks' API monitoring, NeMo Guardrails presents a multi-layered defense against data leaks and sophisticated attacks, while also supporting compliance with data protection laws. Real-world tests have shown that it can boost compliance rates with very little extra delay, helping insurers stay efficient and secure.

The insurance sector's cautious but successful use of AI could set an example for other regulated industries that want to take advantage of this technology without taking on unmanageable risks. It underscores that building reliable AI isn't just about making the model smarter—it's about coordinating safety policies across the whole system and lifecycle. Ongoing monitoring, learning, and skilled human oversight remain essential to keep AI aligned with both laws and ethical values like fairness, clarity, privacy, and accountability.

Looking ahead, there are several important directions for research and future development to make AI in insurance even safer and more effective:

- Improving Hallucination Management: Research should aim for smarter ways to prevent AI from inventing information, perhaps through tighter ties between AI, knowledge graphs, and deeper reasoning techniques for accuracy in complex insurance cases.
- Proactive Bias Handling: Building adaptive guardrails that spot and correct biases in real-time during AI use, rather than only during training, is a future priority.
- Transparent Guardrails: Efforts are needed to make AI guardrail decisions explainable, so users know why outputs are blocked or changed, making audits and reviews easier.
- Shared Safety Standards: Supporting the push for universal safety and fairness measures, as championed by organizations like IEEE, will help insurers reliably compare and improve their AI systems.
- Multi-Agent Collaboration: Investigating how guardrails work in teams of cooperating AI systems will be crucial as insurance applications become more complex.
- Regulatory Adaptation: And finally, there must be ongoing research to make sure technical safety tools keep up with global regulations, making it easier for companies operating in multiple countries to maintain compliance.

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] [1] T. Hartley, "Applications of Artificial Intelligence in Property and Casualty Insurance," *Journal of Risk and Insurance*, vol. 89, no. 2, pp. 254–270, 2022.

[2] Y. Lee and M. Wong, "AI-driven Fraud Detection in Insurance Claims," *IEEE Access*, vol. 9, pp. 123456–123466, 2021.

[3] L. Smith et al., "Automating Customer Service Using NLP in Insurance," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 6, pp. 4901–4908, 2021.
[4] S. Patel and J. Chen, "Evaluating the Effectiveness of Insurance Chatbots," *Insurance: Mathematics and Economics*, vol. 101, pp. 143–150, 2022.
[5] T. Wolf et al., "NVIDIA NeMo Guardrails: Enabling Safe and Secure Conversational AI," in *Proceedings of ICML Workshop on AI Safety*, 2023.
[6] L. Davis et al., "Adversarial Attacks on Insurance Fraud Detection Systems," *Proceedings of the Conference on AI Security*, pp. 67–82, 2024.
[7] S. Miller, "Customer Trust in AI-Powered Insurance Services," *International Journal of Insurance Studies*, vol. 22, no. 1, pp. 12–28, 2024.
[8] Meta AI Research, "Llama Guard: Content Safety for Large Language Models," *Proceedings of the AI Safety Workshop*, pp. 156–171, 2024.
[9] P. Anderson, "Regulatory Compliance in Insurance AI: Challenges and Solutions," *Regulatory Technology Review*, vol. 6, no. 3, pp. 89–106, 2024.

AUTHOR

Rakesh More, Application Support Sr Manager – AI in A. J. Gallagher USA