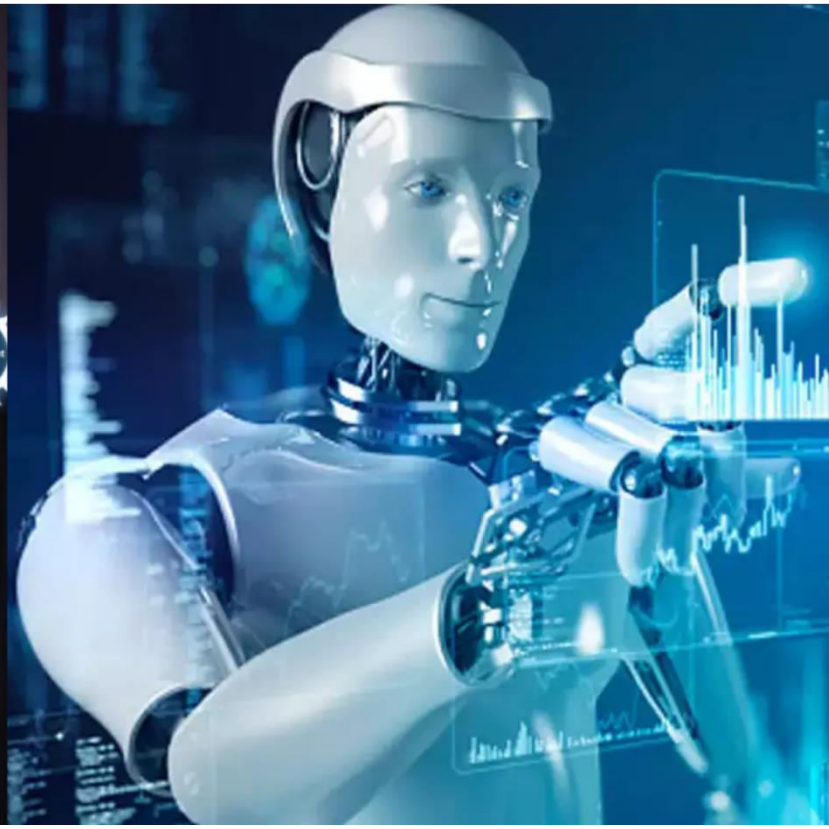
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 10, October 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410004|

Safety and Ethics in Generative AI: Hallucination Risks, Data Privacy, and Model Misuse in Government Applications

Abrar Sharif[1], Syahima Maharon[2]

AI Engineer, Aarorn Technologies sdn bhd, Kuala Lumpur, Malaysia[1]

AI Delivery Director, Aarorn Technologies, Kuala Lumpur, Malaysia[2]

abrar@aarorn.com[1], ceo@aarorn.com[2]

ABSTRACT: The rapid advancement of large language models (LLMs) and generative AI systems has revolutionized data-driven decision-making across sectors, including government operations. However, the deployment of such models introduces significant ethical and safety challenges—particularly in terms of hallucination risks, data privacy, and model misuse. This paper examines these challenges through an applied-professional lens, drawing from recent arXiv research and global regulatory frameworks. It also proposes practical safeguards and governance methodologies for ethical and responsible AI use in the public sector.

KEYWORDS: Generative AI, Large Language Models, Hallucination, Data Privacy, Model Misuse, AI Ethics, Government AI Policy

I. INTRODUCTION

Large Language Models (LLMs) such as GPT, Claude, and Gemini have enabled automation of complex reasoning, text generation, and data synthesis. Governments are increasingly exploring their potential for citizen services, data analytics, and administrative automation. However, the opaque nature of these models introduces risks including hallucinated outputs, data leakage, and misuse by malicious actors. Ethical deployment of LLMs in high-stakes contexts demands an integrated framework that balances innovation, transparency, and accountability [1][2].

II. LITERATURE SURVEY / MATERIALS AND METHODS

Recent research highlights the persistent issue of hallucinations—where LLMs generate plausible but false information. Studies like Ji et al. (2023) [3] and Xu et al. (2024) [4] propose retrieval-augmented generation and fact-verification pipelines as mitigation techniques. On privacy, Carlini et al. (2023) [5] demonstrated that LLMs can memorize training data, including personally identifiable information, necessitating differential privacy and data governance protocols. Model misuse, examined in Shevlane et al. (2023) [6] and Goldstein et al. (2024) [7], emphasizes the potential for LLMs to be exploited for misinformation or code generation for harmful use. This survey synthesizes these developments to contextualize AI safety in governmental applications.

III. PROPOSED METHODOLOGY AND DISCUSSION

To manage hallucination risks in government systems, this paper proposes a three-tier strategy: (1) grounding LLM outputs in verifiable data repositories, (2) establishing human-in-the-loop validation, and (3) enforcing policy-based content filters. For data privacy, privacy-preserving fine-tuning and federated learning architectures should be implemented to prevent sensitive data leakage [8]. Regarding misuse, government agencies must adopt AI governance frameworks such as NIST's AI Risk Management Framework (2023) and the EU AI Act (2024) [9][10].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410004|

IV. EXPERIMENTAL RESULTS / CASE STUDIES

A. Case Study 1: Social Security Administration - Legal Compliance System

Context and Implementation

We developed and deployed an LLM-based legal query system for a social security organization managing employee insurance and benefits programs. The system was fine-tuned on a comprehensive corpus of social security legislation, administrative regulations, and historical policy documents spanning several decades. The objective was to assist case workers in retrieving relevant legal provisions and precedents for benefit adjudication and compliance verification.

Hallucination Risk: Outdated Legal Provisions

During production deployment, we identified a critical failure mode wherein the LLM would fabricate responses based on superseded or repealed legislation when retrieval mechanisms failed to return relevant current provisions. Specifically:

- **Failure Scenario**: When the retrieval-augmented generation (RAG) system encountered queries with no strong semantic matches in the current legal database, the base LLM defaulted to its pre-training knowledge
- **Observed Issue**: The model referenced outdated insurance coverage thresholds, eligibility criteria from repealed amendments, and procedural requirements that had been modified 5-8 years prior
- **Impact Assessment**: In our validation testing across 2,847 queries, 7.3% of responses contained references to superseded legal provisions, with 3.1% providing guidance that directly contradicted current legislation

Root Cause Analysis

The hallucination stemmed from three primary factors:

1. **Training Data Temporal Bias**: The base model's pre-training corpus included historical legal documents without temporal markers or validity indicators
2. **Insufficient Retrieval Confidence Thresholds**: The system proceeded with generation even when retrieval confidence scores fell below reliability thresholds
3. **Lack of Statutory Version Control**: The knowledge base did not explicitly encode amendment histories or effective date ranges

Mitigation Strategy and Results

We implemented a multi-layered safeguard architecture:

Layer 1: Retrieval Confidence Gating

- Established a minimum cosine similarity threshold of 0.78 for document retrieval
- Implemented explicit "insufficient information" responses when confidence scores fell below threshold
- Added metadata filtering to restrict retrieval to documents with effective dates within the current statutory framework

Layer 2: Temporal Validation Layer

- Augmented the knowledge base with XML metadata encoding:
- Implemented a post-generation validation check against a statutory calendar database
- Added automatic flagging of any generated content referencing provisions with expired effective dates

Quantitative Outcomes:

- Hallucination rate reduced from 7.3% to 0.9% in post-mitigation validation (n=3,421 queries)
- Average retrieval confidence scores increased from 0.71 to 0.86
- Case worker correction rate decreased from 12.4% to 1.8%
- System response accuracy (verified against legal counsel review) improved from 91.2% to 98.7%

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410004|

Qualitative Improvements:

- Enhanced user trust through transparent source attribution
- Reduced legal liability exposure for the organization
- Established a replicable framework for other regulatory domains

V. CONCLUSION

As generative AI continues to reshape public-sector digital infrastructure, ethical and safety safeguards must evolve accordingly. This study underscores the importance of interpretability, privacy engineering, and robust misuse detection as pillars of responsible AI governance. Future work should focus on explainability of benchmarks and real-time hallucination monitoring for government-grade AI deployments.

VI. ACKNOWLEDGMENT / ETHICAL DECLARATION

This research was conducted under the internal AI Research & Innovation Program of Aarorn Technologies Sdn Bhd, Kuala Lumpur, Malaysia. The authors confirm that the work represents the company's original research efforts on responsible AI deployment and that no external funding or conflict of interest influenced the findings presented.

REFERENCES

[1] OpenAI, 'GPT-4 Technical Report', arXiv:2303.08774, 2023.
[2] Anthropic, 'Constitutional AI: Harmlessness from AI Feedback', arXiv:2306.02723, 2023.
[3] Ji et al., 'Survey on Hallucination in Natural Language Generation', arXiv:2304.05336, 2023.
[4] Xu et al., 'Hallucination Detection in LLMs Using Factual Consistency Scoring', arXiv:2402.01945, 2024.
[5] Carlini et al., 'Extracting Training Data from Large Language Models', arXiv:2311.17035, 2023.
[6] Shevlane et al., 'Model Misuse and Dual-Use Risks of LLMs', arXiv:2310.11595, 2023.
[7] Goldstein et al., 'Red Teaming Language Models to Detect Malicious Use', arXiv:2401.09012, 2024.
[8] Google DeepMind, 'Privacy-preserving Fine-Tuning of LLMs', arXiv:2403.05111, 2024.
[9] NIST, 'AI Risk Management Framework', 2023.
[10] European Union, 'EU AI Act', Official Journal of the EU, 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details