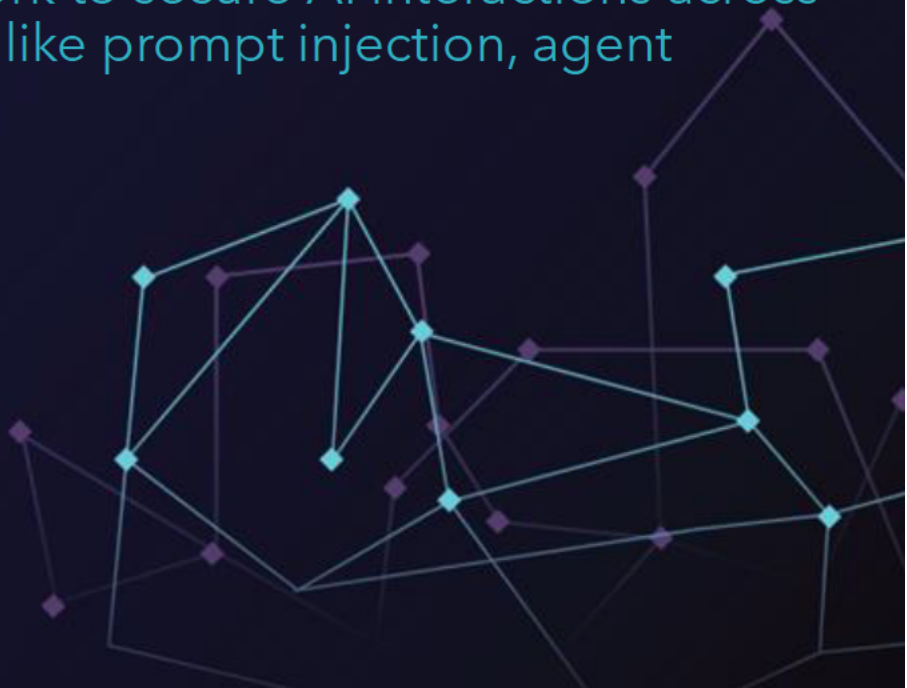
Australian Government | GovAI

GovTrust AI: Empowering Secure and Safe Adoption of AI

GovTrust AI provides a comprehensive framework to secure AI interactions across government workflows, addressing critical risks like prompt injection, agent misuse, and opaque model outputs.



CSIRO



Australian Government | GovAI



Executive Summary

Problem Faced	Solution Overview	Benefits and Impact
CSIRO and other government agencies are using AI for drafting, summarization and decision support, creating new systemic risks beyond traditional cybersecurity. • Prompt Fragility: Inputs to LLMs can be subtly manipulated. • Agentic Misbehaviour: AI agents may take unintended or unauthorized actions. • Adversarial Manipulation: Malicious actors can exploit model behavior via prompt injection or adversarial prompts. • Lack of System-Level Safeguards: Most security efforts focus on model robustness.	CSIRO is developing secure and responsible AI for high-assurance environments, with a focus on validation, monitoring, and system-level safety. The approach is broadly adaptable. Key components: • Adversarial Taxonomy: Based on our survey, mapping manipulation types and system impacts. • LLM Red Teaming: Evaluates model behavior under adversarial conditions, focusing on input/output security. • Secure I/O Handling: Real-time input validation and anomaly detection to block prompt injection and unsafe outputs. • System Oversight: Ongoing evaluation for bias, robustness, misuse, and compliance.	CSIRO's GovTrustAI framework addresses emerging risks from AI use in government decision-making by mitigating prompt fragility, agentic misbehaviour, and adversarial manipulation. By embedding system-level safeguards beyond model robustness, it enhances resilience, ensures auditable and compliant outputs, enables early risk detection, and supports cross-sector AI safety and policy standards.



Target Audience and Stakeholders

- Government agencies adopting AI in critical workflows
- AI system developers and integrators
- Cybersecurity and risk management teams
- Regulators and policy makers focused on AI governance
- Researchers in AI safety and adversarial robustness
- Enterprises deploying AI in high-assurance environments

Risks and Mitigation Overview

Over-Dependence: Excessive AI reliance
Mitigation: Human-in-loop reviews & continuous monitoring
Misinterpretation: Misuse of AI outputs
Mitigation: Explainability tools & user training
Adversarial Manipulation: Malicious prompts
Mitigation: Secure I/O handling & red teaming
Agentic Misbehaviour: Unauthorized AI actions
Mitigation: Model oversight & system-level security-by-design

Use Case Status

In Development

Use case timeline

- September 2024- Feburary 2025: Literature Review, Taxonomy mapping and identification of adversarial misuse.
- Feburary 2025- Ongoing: RedTeaming for Secure input and output handling.
- Model evaluation: In progress
- Depending on funding: System security



Additional Information

Data Considerations

- Prompt Logs: Inputs submitted by users during interactions with AI tools.
- Model Responses: AI-generated outputs, summaries, reports, or recommendations.
- Interaction Metadata: Timestamps, user roles, model versions, and tool usage patterns.
- Policy Reference Data : Publicly available or internal policy documents used to benchmark output compliance.

Lessons Learned

Objective-based adversarial taxonomy enables more effective threat detection and defense.

Robust red teaming, auditing, and monitoring are essential for continuous model oversight.

Real-time I/O validation is critical to prevent prompt injection and unsafe outputs.

System-level, safety-by-design approaches are needed to secure all model components.

Adaptive and privacy-preserving defenses are vital for resilient, high-assurance AI deployment.

Contact information

Responsible Entity Name

CSIRO's Data61

Open for Collaboration?

Interested in collaboration with other Commonwealth entities

Area of Entity

Software & Computational Systems

Use Case Contact

Sharif Abuadbba:
Sharif.Abuadbba@data61.csiro.au

Use Case Website/s

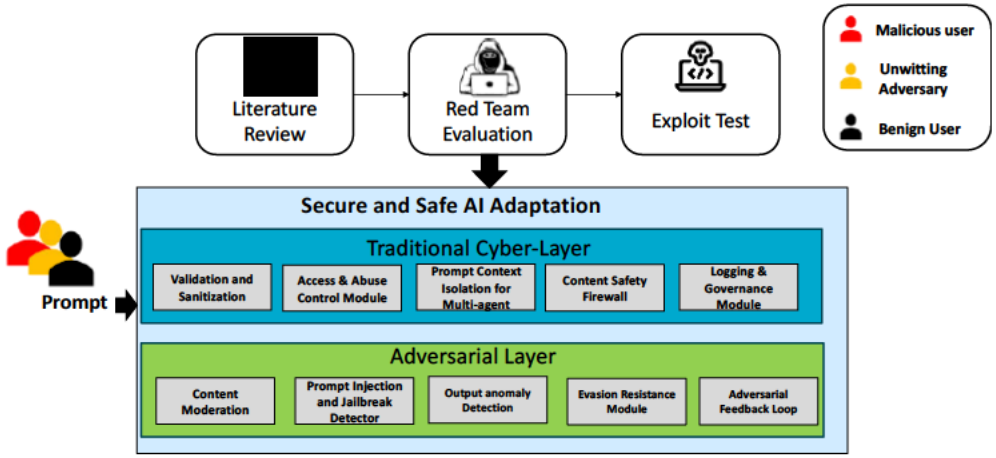
N/A

Use Case Owner

Liming Zhu:
Liming.Zhu@data61.csiro.au

Screenshot/s

Overview of the Usecase



ScreenShot of the Literature Review

private and publicly available data, to develop a general understanding of language patterns, syntax, and semantics. This large-scale pre-training helps the LLM learn to predict the next word in a sequence, forming the backbone, namely foundation model, of their language generation capability. Following pre-training, the LLM foundation model can undergo a fine-tuning phase, where they are further trained on more specific, often domain-targeted datasets. This additional training allows the models to specialize in particular tasks or areas of expertise, adapting their learned language capabilities to meet the needs of specific applications, such as question answering, summarization, or translation.

Once deployed, users can engage with LLMs through an intermediary agent, submitting queries that initiate a dynamic interaction. The agent communicates with the LLM, potentially enhancing its responses through the use of a Retrieval-Augmented Generation (RAG) system, which integrates external knowledge retrieval to improve answer accuracy and relevance. Throughout this process, the interaction is safeguarded by comprehensive guardrail systems, which monitor, filter, and manage inputs and outputs to ensure compliance with safety, ethical, and operational standards. These guardrails play a crucial role in screening queries and responses, ensuring that the information returned to the user aligns with established guidelines for responsible AI use.

This architecture highlights the intricate flow of data and decisions within LLM system interactions, emphasizing the interplay between core model components, external data augmentation, and protective measures to maintain secure and effective communication.

2.2. Taxonomy Overview

Figure 2 provides an overview of the four categories of LLM attacks, namely *privacy breach*, *integrity compromise*, *availability disruption* and *misuse*. Throughout the paper, we will systematically explore representative attackers attempt to make alterations on such four key components to change the behaviours or extract secret information from LLMs.

2.4. Techniques

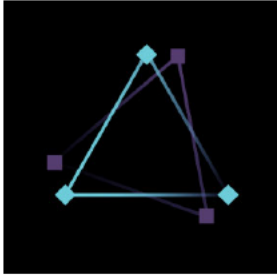


Figure 2: The overview of the four categories of LLM attacks.

Attackers tend to leverage various techniques to mount attacks. We begin with a brief introduction to these general techniques.

- **Data Poisoning** aims to intentionally inject misleading or malicious data into the training dataset of a model with the goal of causing the model to make incorrect predictions or behave in unintended ways.
- **Weight Poisoning** modifies the parameters (weights) of a models by manipulating the training algorithm or by having access to the model updates.
- **Model Editing** refers to the process of making targeted knowledge changes or updates to the behavior of a pre-trained model by (surgically) modifying its parameters, structure, or by adding specific components.
- **Prompt Injection** involves inserting malicious or deceptive prompts into the input that can manipulate the

Available online: https://arxiv.org/pdf/2502.02960



Detailed Overview

Version Control

Version	Date	Author	Description of Changes
1.0	3 Feb 2025	GovAI	Version 1 created
1.1	17 Mar 2025	GovAI	Modified based on feedback

Index

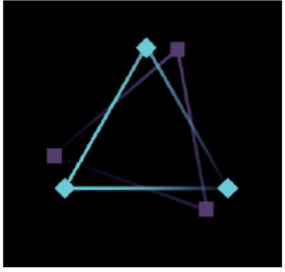
Responsible Organisation Category 5
Scope of the Use Case 6
Ethical Considerations 6
Value of the Use Case 7
AI Process Type 8
AI Technologies Utilised 9
Technical Elements 10

Note: For details about category items in the detailed overview, see APS AI Use Case Repository Guidance- Guidance for Use Case Owners and Editors.

Responsible Organisation Category

Select the Classification of the Functions of Government – Australia (COFOG-A) 3-digit category that best identifies the functional area associated with your AI use case.

☒ **01 - General Public Services**	019 - General public services (other)
☐ **02 - Defence**	Choose an item.
☒ **03 - Public Order and Safety**	035 - Research and development in public order and safety
☒ **04 - Economic Affairs**	045 - Communication
☐ **05 - Environmental Protection**	Choose an item.
☐ **06 - Housing and Community Amenities**	Choose an item.
☐ **07 - Health**	Choose an item.
☒ **08 - Recreation, Culture, and Religion**	085 - Research and development in recreation, culture, and religion
☒ **09 - Education**	096 - Research and development in education
☒ **10 - Social Protection**	107 - Research and development in social protection
☒ **11 - Transport**	118 - Research and development in transport



Scope of the Use Case

Use the dropdown menus below to identify the scope of your use case.

Geographical focus Choose the region for implementation from the dropdown list	National
Primary type of government interaction Choose the type of government interaction from the dropdown list	Government-to-government (G2G)
Cross-features - Sector Indicate if the use case describes a solution that can be used across sectors or in cross-sector scenarios (Yes/No).	No
Cross-features – Jurisdiction Indicate if the use case describes a solution that can be used across State/Federal borders or in cross-border scenarios (Yes/No)	Yes

Ethical Considerations

Accuracy, Fairness, Accessibility, Bias and Discrimination	• GovTrustAI will implement multi-layered validation processes, including human-in-the-loop oversight and policy-aware content review, to ensure that AI-generated outputs are accurate, representative, and aligned with government standards. • Bias detection mechanisms will be incorporated to audit model behaviour and minimize discriminatory outcomes. • The system will be designed to support accessibility for diverse user groups and ensure inclusive AI interactions.



Privacy	• GovTrustAI will avoid processing personally identifiable information unless explicitly authorized and governed by appropriate legal frameworks. • Sensitive data will be encrypted and stored off-chain to maintain confidentiality and compliance with privacy regulations.
Rights of Users	• Users will be clearly informed about how their interactions with AI are processed and logged. • The platform will include mechanisms for users to provide feedback, flag concerns, and challenge AI-assisted outputs. Transparency features and audit tools will support user awareness and enable redress or escalation if necessary.

Value of the Use Case

Identify the public value that the solution provides or is expected to provide. Select from the multi-select options.

Improved public service This category refers to solutions that enhance the services provided to end users, whether they are citizens or businesses.	☒ Personalised services ☐ Public (citizen)-centred services ☒ Increased quality of public information and services ☒ More responsive, efficient and cost-effective public services ☐ New services or channels
Improved administrative efficiency This category refers to solutions that increase efficiency, effectiveness, and quality while reducing costs within administrative processes, systems, and services.	☐ Cost reduction ☐ Responsiveness of government operation ☒ Improved management of public resources ☒ Increased quality of processes and systems ☒ Better collaboration and better communication



	☐ Reduced risk of corruption and abuse of the law by public servants ☐ Greater fairness, honesty and equality enabled
Open government capabilities This category refers to solutions that enhance the level of openness, transparency, engagement, and communication within public organisations.	☐ Increased transparency of public sector operations ☒ Increased public participation in government actions and policymaking ☒ Improved public control of and influence on government actions and policies

AI Process Type

Select the types of tasks within government operations that the AI solution is performing or expected to perform

Supporting Decision Making- Tasks that support formal or informal agency decision-making on benefits or rights.	☒ Taking decisions on benefits ☒ Managing copyright and intellectual property rights
Analysis, monitoring and regulatory research – Tasks that collect or analyse information that shapes agency policymaking.	☒ Information analysis processes ☒ Monitoring policy implementation ☐ Innovating public policy ☒ Prediction and planning
Enforcement - Tasks that identify or prioritise targets of agency enforcement action.	☐ Smart recognition processes ☒ Management of auditing and logging ☐ Predictive enforcement processes ☐ Supporting inspection processes ☒ Improving cybersecurity ☐ Registration and data notarisation processes ☒ Certification and validation processes
Internal management - Tasks that support agency management of resources, including employee management, procurement, and maintenance of technology systems.	☒ Internal primary processes ☒ Internal support processes ☒ Internal management processes ☒ Procurement management ☒ Financial management and support
Public services and engagement - Tasks that support the direct provision of services to the public or facilitate	☒ Engagement management ☒ Data-sharing management ☐ Governance and voting



communication with the public for regulatory or other purposes.	☐ Payments and international transactions ☐ Supporting disintermediation ☐ Authentication of self-sovereign digital ID services ☐ Service integration ☐ Service personalisation ☐ Tracking of goods and assets along the supply chain

AI Technologies Utilised

Select the types of AI technologies proposed / utilised to deliver the use case.

Reasoning or Knowledge Representation AI systems that store, structure, and process knowledge to make inferences, derive conclusions, or support decision-making.	☒ Knowledge Representation ☒ Automated Reasoning ☐ Commonsense Reasoning
Planning and Optimisation AI techniques that generate, refine, and optimise action sequences or resource allocation to achieve specific goals efficiently.	☒ Planning and Scheduling ☐ Searching ☒ Optimisation
Learning and Adaptation AI systems that identify patterns, extract insights, and improve performance over time based on data.	☐ Machine Learning ☒ Deep Learning ☒ Generative AI
Communication and Natural Language Processing AI systems that process, interpret, and generate human language for interaction, comprehension, and automation.	☒ Natural Language Processing (NLP) ☒ Text Generation ☒ Text Mining ☐ Machine Translation
Perception through the Senses AI systems that process and interpret sensory data, such as visual, auditory, or	☒ Computer Vision ☒ Audio Processing



tactile inputs, to understand and respond to their environment.	
Integration and Interaction with the Environment AI systems that interact with physical or digital environments, including autonomous agents, robotics, and interconnected systems.	☒ Multi-agent Systems ☐ Robotics and Automation ☐ Connected and Automated Vehicles (CAVs)
AI as a Service AI capabilities delivered through cloud-based platforms, offering tools, models, and infrastructure for AI-powered applications.	☒ AI Services (e.g., cognitive computing, machine learning frameworks, bots) ☐ Infrastructure as a Service (IaaS) ☐ Platform as a Service (PaaS) ☐ Software as a Service (SaaS)
Additional Comments or Explanation:	If you have selected any of the subcategories above, feel free to provide more detailed comments or a description of how these elements apply to your specific use case.

Technical Elements

Platform implementation	Designed to be deployed on secure cloud platforms like AWS, Azure, or GCP, enabling scalability and high availability. On-premises deployment: For organizations requiring strict data governance, an on-premises deployment option is supported.
Model / Algorithm used	• Large Language Models (LLMs): Pretrained general-purpose LLMs (e.g., GPT, , or government-favored open-source alternatives) will be used for tasks such as summarization, content generation, and classification. These models will be monitored and scoped for safe use within predefined policy boundaries. • Input Sanitization Models: Lightweight NLP models will detect



	and filter potentially harmful or adversarial inputs in real time to prevent prompt injection and ensure safe interaction. • Output Validation Models: AI classifiers and rule-based engines will evaluate AI-generated content for policy compliance, bias detection, misinformation, or hallucination risks. This includes keyword filtering, semantic checks, and policy alignment scoring. • Audit & Explainability Algorithms: Attribution models and explainability tools (e.g., attention visualization, SHAP, or LIME-like techniques) will be used to trace how outputs were generated and why specific content was included. • Blockchain Smart Contracts:Custom smart contracts will be implemented to record key interaction metadata (e.g., model version, input type, output action) in a tamper-proof, timestamped format for audit and traceability.	
Data Sources Select the types of data sources used and provide relevant details.	☒ Internal ☒ Public	☒ Third-party ☒ Synthetic
	Details: Training will be done on the hybrid of internal, public, third-party and synthetic data. As AI-driven integrated risk management system will manage risk across department that include	
Risk Assessment and Mitigation Details		
Security and Compliance Frameworks Select the security and compliance frameworks and measures	☐ Authority to Operate (ATO) ☒ System Security Plan (SSP)	☐ Information Security Registered Assessors Program (IRAP)



implemented. Provide details or additional artifacts if relevant.	☒ Security Risk Management Plan (SRMP)	☐ Penetration Testing
	Details:	
Assurance and Government Frameworks	Identify the assurance or government frameworks you have evaluated your use case against so far (e.g., DTA AI Assurance Framework, AI in Government Policy, National AI Assurance Framework).	
Record maintenance	Provide an overview of documentation practices for AI decisions, testing, and data assets.	
Disengagement	Describe any contingency plans for disengagement in the event of critical failures (N/A if in early planning stage)	
Performance Metrics and Results	Quantitative Metrics • Model Accuracy: Percentage of correct responses generated by AI models, aiming for 95%+ accuracy. • Prompt Injection Mitigation: Rate at which prompt injections and adversarial inputs are detected and prevented, targeting 98%+ resistance. • Blockchain Integrity: Percentage of AI interactions logged on blockchain with 100% audit trail completeness and zero tampering. • Output Compliance: Percentage of AI-generated content meeting policy compliance, aiming for 98%+ compliance rate. • System Efficiency: Average response time under 1 second and throughput of 1000+ queries per minute. • User Trust: Percentage of users who trust AI-generated outputs, aiming for 85%+ trust. Qualitative Metrics • User Satisfaction: Feedback-based score measuring overall user	



	confidence and contentment with the system, aiming for 90%+ satisfaction. • Model Explainability: Transparency of AI decision-making and auditability of outputs for accountability, with clear reasoning behind AI actions. • Ethical Compliance: Ensuring AI outputs are free from bias, compliant with legal standards, and ethically sound. • System Drift Monitoring: Monitoring system performance for drift or degradation, ensuring timely retraining to maintain accuracy and reliability.