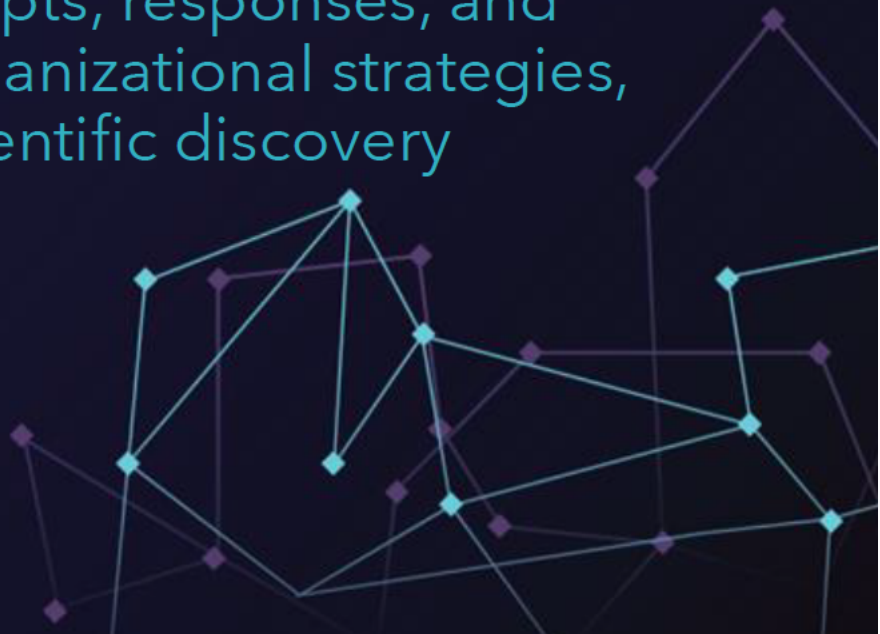
Australian Government | GovAI

AI Guardrails Assistant

AI-enabled assistant that evaluates prompts, responses, and documents to ensure alignment with organizational strategies, policies and quality management for scientific discovery



CSIRO

Australian Government | GovAI



Executive Summary

Problem Faced

CSIRO faces key risks when using AI tools to process scientific data or documents with protected information (e.g., intellectual property (IP) and/or confidential/private information):

- Unintentional scientific information leaks through prompts or AI responses.
- Need to meet privacy/legal standards.
- Ensuring AI-generated content aligns with CSIRO's privacy/IP policies and government strategies.
- Lack of tools to review and revise documents for IP/privacy and general policy alignment and quality assurance.

Solution Overview

The AI Guardrail Assistant offers a large-language-model-based tool to detect protected information and support compliance with legal, regulatory, and ethical standards. It provides the following features that cannot be performed by the conventional IT solutions.

- Detection of potential information leaks in scientific prompts and documents involving protected information.
- Automatic output checks for alignment with policies and strategies.
- Support for user-defined criteria.
- Risk and quality management to reduce legal, reputational, and operational issues while maintaining consistency and document quality.

Benefits and Impact

- Efficiency Gains – Automates checks for information leaks, policy compliance, and risks, reducing manual review time.
- Cost Savings – Lowers legal and compliance costs by preventing breaches and policy violations.
- Improved Decision-Making – Supports transparency, policy adherence, and trust in AI use for scientific discovery.
- Broader Impacts – Could be extended to other industrial and government use cases to ensure AI output alignment with organizational strategies, policies and quality management



Target Audience and Stakeholders

- Scientists and researchers
- Information governance and ethics officers
- Policy and compliance teams
- Executive and business unit leadership
- Program and project managers

Risks and Mitigation Overview

Privacy and confidentiality risks

- Risk: AI may inadvertently expose confidential or proprietary data.
- Mitigation: Use confidential data detection and automated compliance checks to prevent leakage.

Regulatory and compliance challenges

- Risk: AI outputs may not meet legal, ethical, or sector-specific requirements.
- Mitigation: Regular updates of user-defined criteria to reflect evolving regulations and automated policy alignment.

Use Case Status

Implemented

Use case timeline

Feb 2025: Prototype developed with an initial focus on privacy and IP information protection for the digital science of plant genomics
May/Jun 2025: Limited internal deployment
July 2025: Stakeholder feedback collection
August 2025: Iterative improvements and expending scopes on funding availability



Additional Information

Data Considerations

- Types of Data: Includes confidential data, IP content, organisational policies, documents, and user inputs.
- Data Governance: Ensures compliance with legal and regulatory standards through policy summarisation and alignment with established guidelines.

Lessons Learned

- Ensuring AI compliance required detailed mapping of diverse legal, regulatory, and ethical policies.
- A user study is essential to define and contextualise what constitutes sensitive data.

Contact information

Responsible Entity Name

CSIRO's Data61

Area of Entity

Software & Computational Systems

Use Case Website/s

https://research.csiro.au/isp/research/privacy_mlai/managing-confidential-data-and-transparency-policies-in-llm-empowered-science/

Open for Collaboration?

Interested in collaboration with other Commonwealth entities

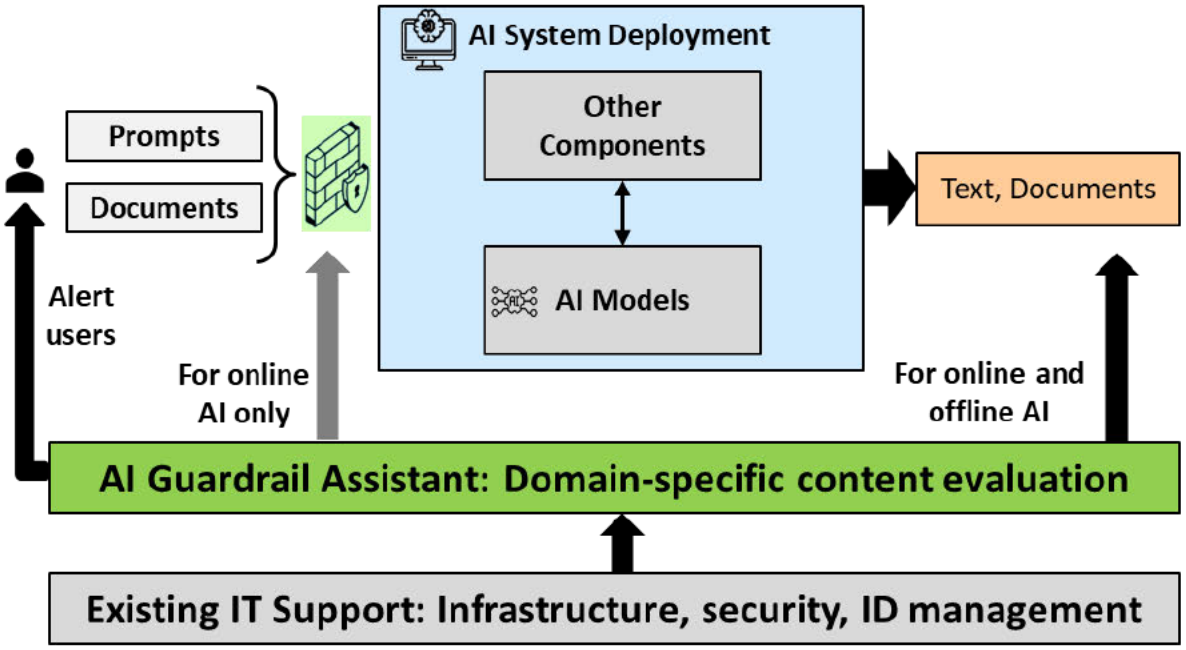
Use Case Contact

Ming Ding: Ming.Ding@data61.csiro.au
Qinghua Lu: Qinghua.Lu@data61.csiro.au

Use Case Owner

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

Screenshot/s



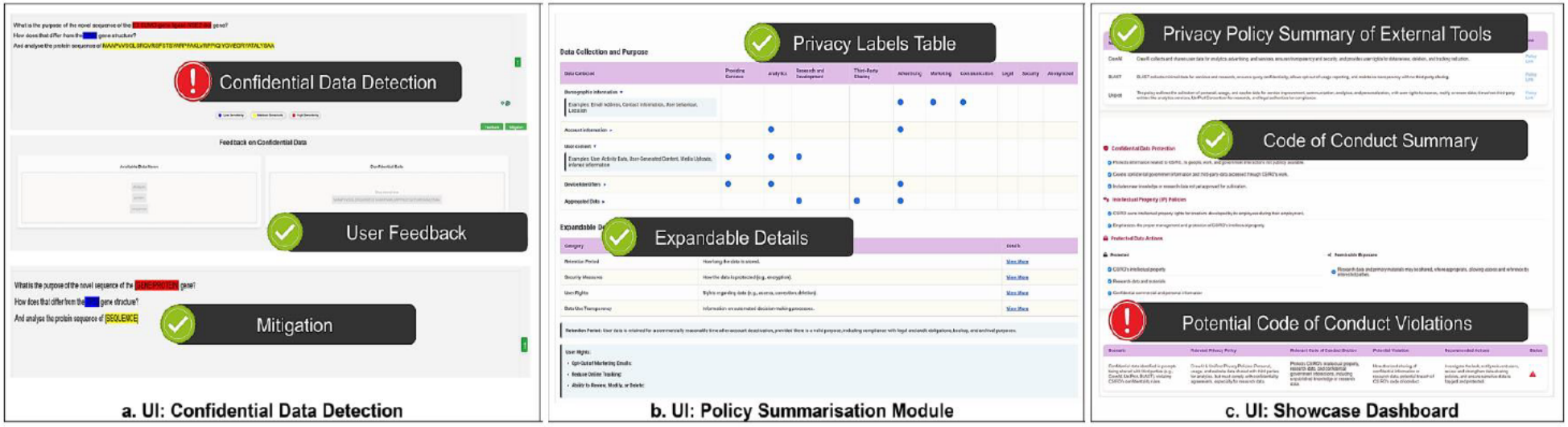
Value Proposition

1. Detects potential leaks of **protected information** in prompts and documents
2. Ensures **alignment** with agency policies and strategies
3. Supports **user-defined goals**
4. Manages risk and quality when using AI-generated text

Technical Elements

1. Graph-based policy extraction
2. Large language models
3. Fuzzy matching and rule-based pattern recognition
4. Retrieval-augmented generation for customisation

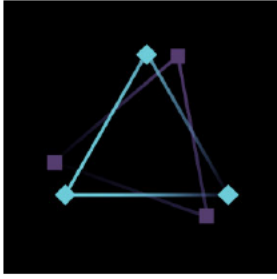
Screenshot/s



a. UI: Confidential Data Detection

b. UI: Policy Summarisation Module

c. UI: Showcase Dashboard



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 6
Scope of the Use Case 7
Ethical Considerations 7
Value of the Use Case 8
AI Process Type 8
AI Technologies Utilised 9
Technical Elements 11

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**	013 - General personnel services
☐ **02 - Defence**	Choose an item.
☒ **03 - Public Order and Safety**	035 - Research and development in public order and safety
☐ **04 - Economic Affairs**	Choose an item.
☒ **05 - Environmental Protection**	055 - Research and development in environmental protection
☒ **06 - Housing and Community Amenities**	065 - Research and development in housing and community amenities
☒ **07 - Health**	077 - Research and development in health
☒ **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	Across countries
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).	Yes
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	We ensure accuracy through trusted policy sources and real-world validation. Fairness is supported by diverse data and regular bias checks. The system is accessible, reflects stakeholder input, and applies policies consistently to prevent discrimination.
Privacy	No personal data will be stored or shared, and all processing follows relevant legal and regulatory requirements, e.g., Australian Privacy Act 1988.
Rights of Users	Feedback mechanisms are built into the system, allowing users to report issues or provide input. Users can challenge AI



	decisions through designated review channels involving human oversight.

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-	☐ Taking decisions on benefits



Tasks that support formal or informal agency decision-making on benefits or rights.	☒ 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 communication with the public for regulatory or other purposes.	☒ Engagement management ☐ Data-sharing management ☐ Governance and voting ☐ 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 tactile inputs, to understand and respond to their environment.	☐ Computer Vision ☐ Audio Processing
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

	1.Hosting platform: Cloud-based deployment: Our developed tool, DataShield, is designed to be deployed on secure cloud platforms like AWS, Azure, or Google Cloud Computing. On-premises deployment: For organizations requiring strict data governance, an on-premises deployment option is supported. 2. Security considerations: We use local language model and retrieval-augmented generation (RAG) to ensure that sensitive data is not stored or leaked during interactions. Moreover, we implement access control mechanisms, including role-based access control. 3. Integrations with existing platforms and technology: We use existing LLM models, tools, and government and organizational policies in our framework to process and assess data transparency and compliance. The solution can seamlessly integrate with existing LLM-based AI systems, research platforms, and laboratory data management systems. 4.Cost model: Since the solution relies on LLM API calls, API-based billing must be considered based on usage.
Model / Algorithm used	LLMs for Confidential Data Detection: • Utilized pre-trained models such as open-source local models like Claude Opus, Claude Sonnet and LLaMA. • Incorporated Retrieval-Augmented Generation (RAG) to enhance model responses by integrating domain-



	specific knowledge from external databases. • Applied fuzzy matching and rule-based pattern recognition for identifying protected information such as sequences and metadata. Privacy Policy Summarization: • Utilized graph-based analysis (PoliGraph) for structured policy extraction. • Used LLM-powered RAG to refine extracted content and generate human-readable summaries. Compliance & Risk Assessment: • Used Natural Language Processing (NLP) and LLM techniques to compare privacy policies against organizational guidelines. • Use LLM pre-trained models to identify violations and misalignment with compliance standards.
Data Sources Select the types of data sources used and provide relevant details.	☐ Internal ☐ Third-party ☒ Public ☒ Synthetic **Details:** We trained the framework using a public dataset and tested it with synthetic prompts.
Risk Assessment and Mitigation Details	Challenges: • Risk of data leakage when using cloud computing services. • Model hallucinations or misinterpretation of protected information. Mitigations: • Utilizing local open-sourced LLMs instead of online models to reduce data exposure to external tools. • Implemented prompt sanitization by applying pattern-based filters and entity redaction (e.g., replacing confidential data with placeholders).



	• Applied human-in-the-loop validation to verify and adjust model predictions.
Security and Compliance Frameworks Select the security and compliance frameworks and measures implemented. Provide details or additional artifacts if relevant.	☐ Authority to Operate (ATO) ☐ System Security Plan (SSP) ☐ Security Risk Management Plan (SRMP) ☐ Information Security Registered Assessors Program (IRAP) ☐ 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: • Confidential Data Detection Accuracy: Precision, recall, and F1-score for identifying confidential data. • Response Time: Latency for processing prompts and generating insights. • Accuracy in policy summarization via question and answering • Compliance Violation Rate: Effectiveness in identifying policy violations. Qualitative Metrics: • User Satisfaction: Based on user study and feedback.