Google Cloud

Protecting the future of AI

How Google Cloud helps mitigating LLM risks and promote a responsible adoption

April 7th 2025

The development of AI is creating **new opportunities** to improve the lives of people around the world, from business to healthcare to education; it's also raising many **questions...**

Regulatory

Safety

Ethics

Privacy

Compliance

Explainability

Governance

THE HILL

AI ‘wild west’ raises national security concerns

WORLD ECONOMIC FORUM

Cybersecurity | Cybercrime

Prompt injection attacks threaten AI chatbots, and other cybersecurity news to know this month

Venture in Security

We have already failed to secure AI by doing what we did before - repeating the mistakes of the past

A different take on the problem of AI security

ROSS HALELIUK
SEP 5, 2023

ZDNET

Trust in ChatGPT is wavering amid plagiarism and security concerns

A new report reveals users' biggest qualms about the popular AI chatbot.

AI may compromise our personal information if companies aren't held responsible

ZDNET

Why AI fails spectacularly at cybersecurity

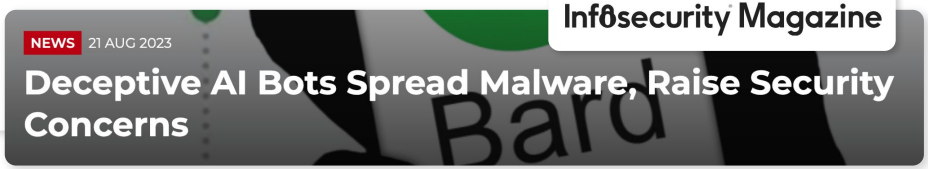
CYDEF

Survey reveals mass concern over generative AI security risks

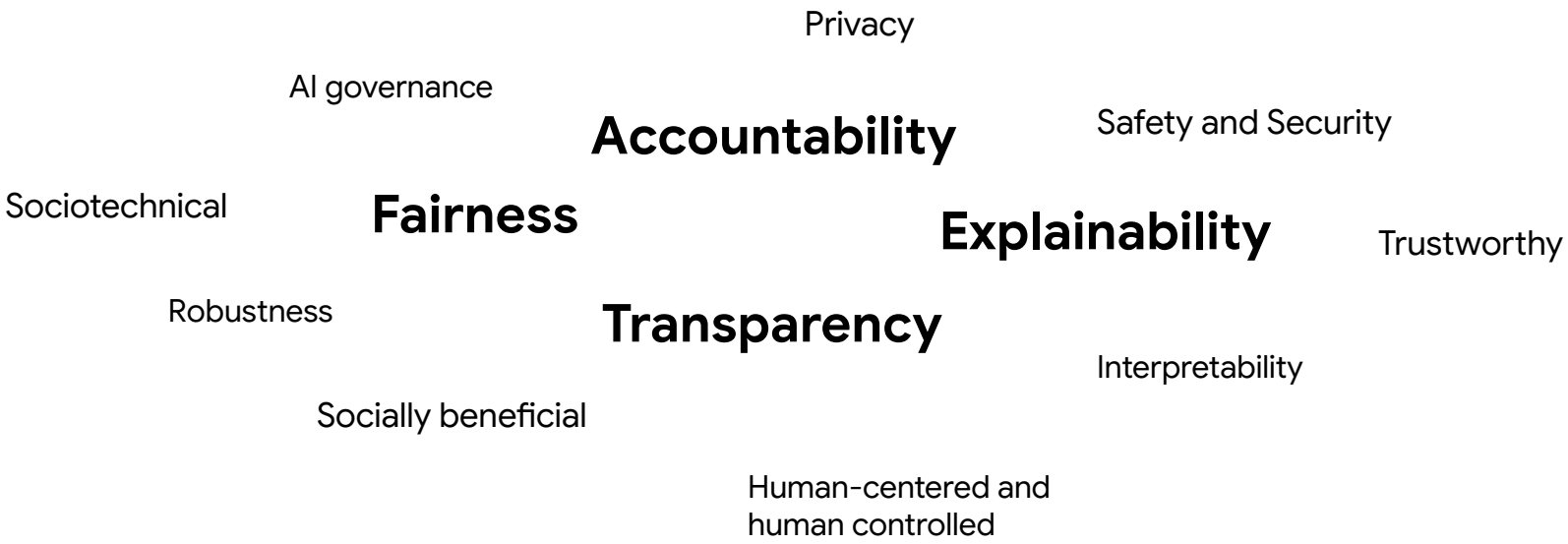
News
Jun 27, 2023 • 3 mins

CSO

Infosecurity Magazine



Responsible AI is



Our work is guided by the AI Principles

AI should:

1. Be socially beneficial
2. Avoid creating or reinforcing unfair bias
3. Be built and tested for safety
4. Be accountable to people
5. Incorporate privacy design principles
6. Uphold high standards of scientific excellence
7. Be made available for uses that accord with these principles

Applications we will not pursue:

1. Likely to cause overall harm
2. Technologies primarily intended to cause injury
3. Surveillance violating internationally accepted norms
4. Purpose contravenes international law and human rights

https://ai.google/principles/

AI Governance Committee

Board of Directors

Risk Management Committee

- IT Risk Committee
- AI Governance Committee
- Vendor Risk Management Committee
- Cyber Risk Committee
- Operational Risk Committee

Proposed Members:

- IT Infrastructure
- Information Security
- Application Security
- Risk
- Compliance
- Privacy
- Legal
- AI Product POC
- Data Protection / Governance
- Third Party Risk Management
- Other experts, as needed *(e.g., HR)*

Three ways we build security into the Google Cloud

Security by design

- Multiple levels of encryption plus physical and logical security
- Comprehensive security throughout the stack
- Rigorous controls of our software supply chain and binary authorization
- Security testing, red-teaming, threat analysis, and vulnerability research

Security by default

- Continuously updated security defaults
- Security blueprints
- Embedded default security configurations
- Default encryption at rest

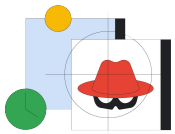
Security in deployment

- Security built on experience
- Tools and monitoring to provide continuous assurance
- Encryption at rest and in transit
- Advanced security features

Google's Secure AI Framework (SAIF)

AI is advancing rapidly, and it's important that
effective risk management strategies evolve along with it

Expand strong security foundations to the AI ecosystem



Extend detection and response to bring AI into an organization's threat universe

Automate defenses to keep pace with existing and new threats

Harmonize platform level controls to ensure consistent security across the organization

Adapt controls to adjust mitigations and create faster feedback loops for AI deployment

Contextualize AI system risks in surrounding business processes

https://safety.google/cybersecurity-advancements/saif/

SAIF Risk Map

https://saif.google

Application

1. Denial of ML Service
2. Insecure Integrated Component
3. Model Reverse Engineering
4. Rogue actions

Model

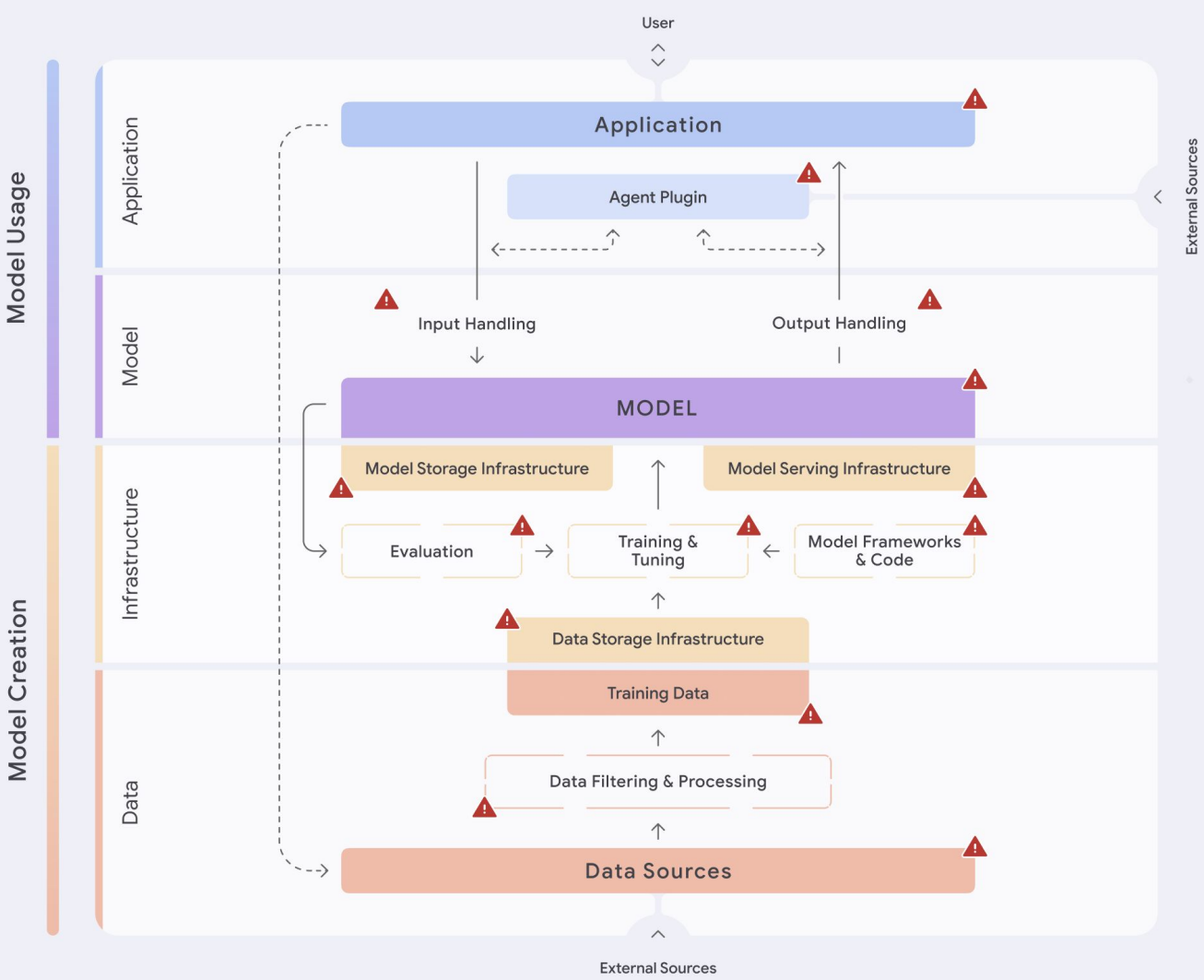
5. Sensitive Data Disclosure
6. Inferred Sensitive Information
7. Prompt Injection
8. Model Evasion
9. Insecure Model Output

Infrastructure

10. Excessive Data Handling
11. Model Source Tampering
12. Model Exfiltration
13. Model Deployment Tampering

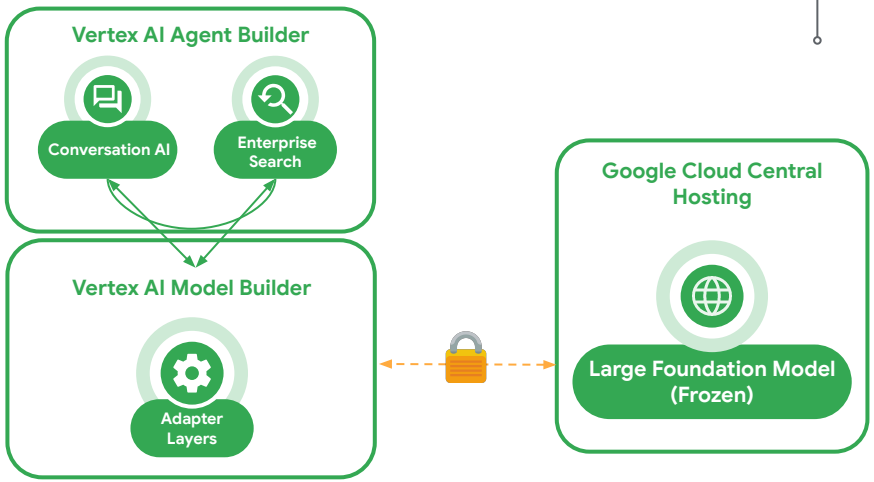
Data

14. Data Poisoning
15. Unauthorized Training Data

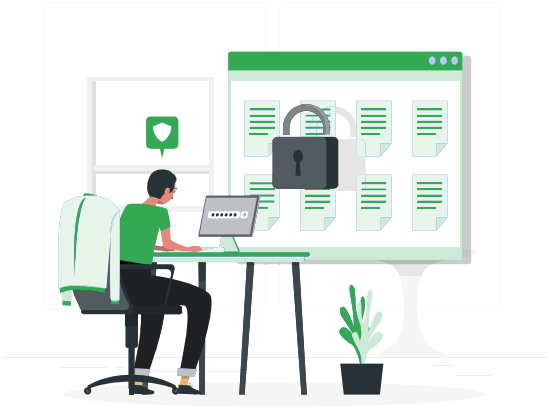


Model Tuning

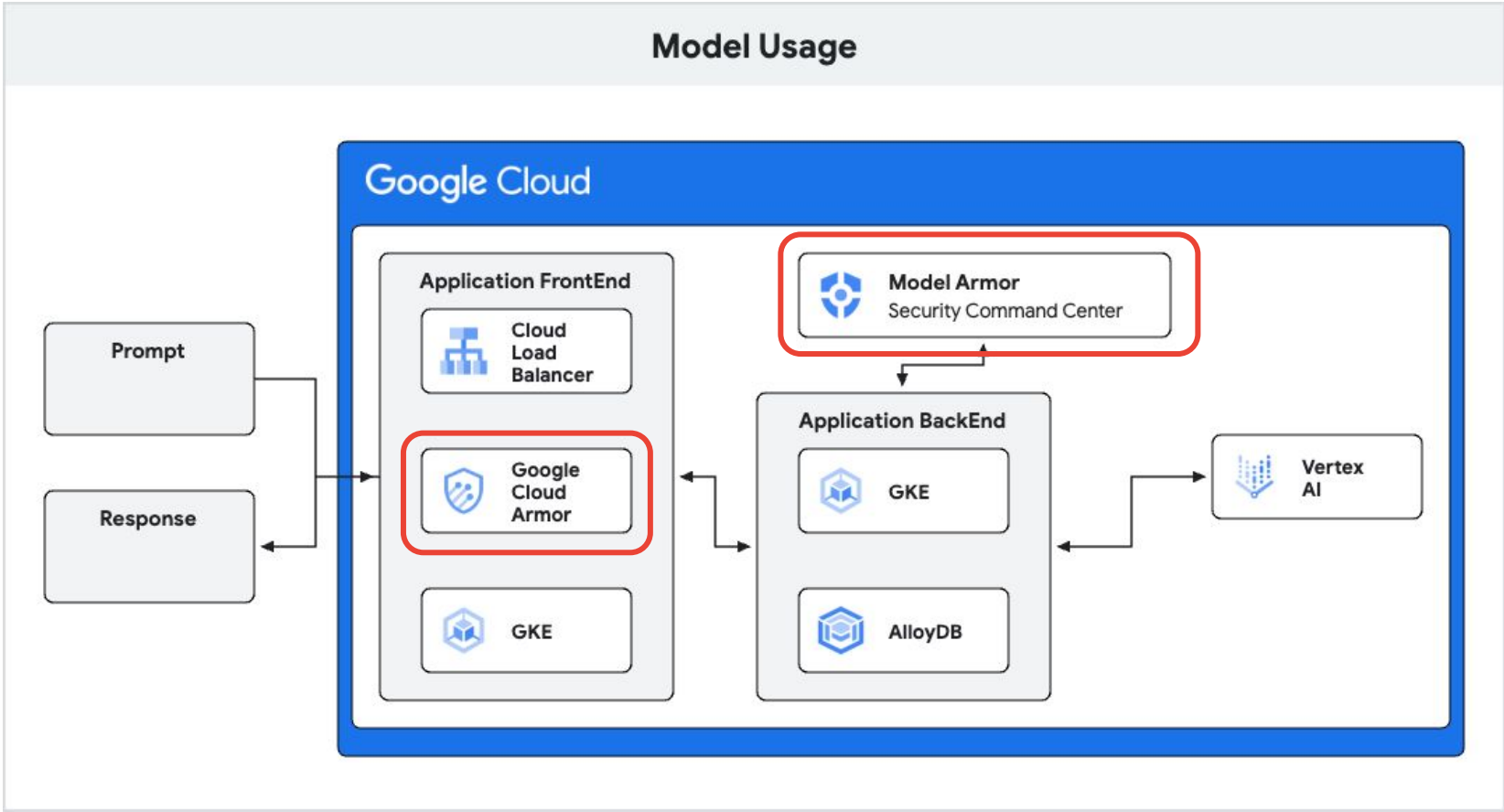
- Parameter Efficient Fine Tuning (PEFT)
- Customer specific adapter weights
- Foundational model remains frozen during inference



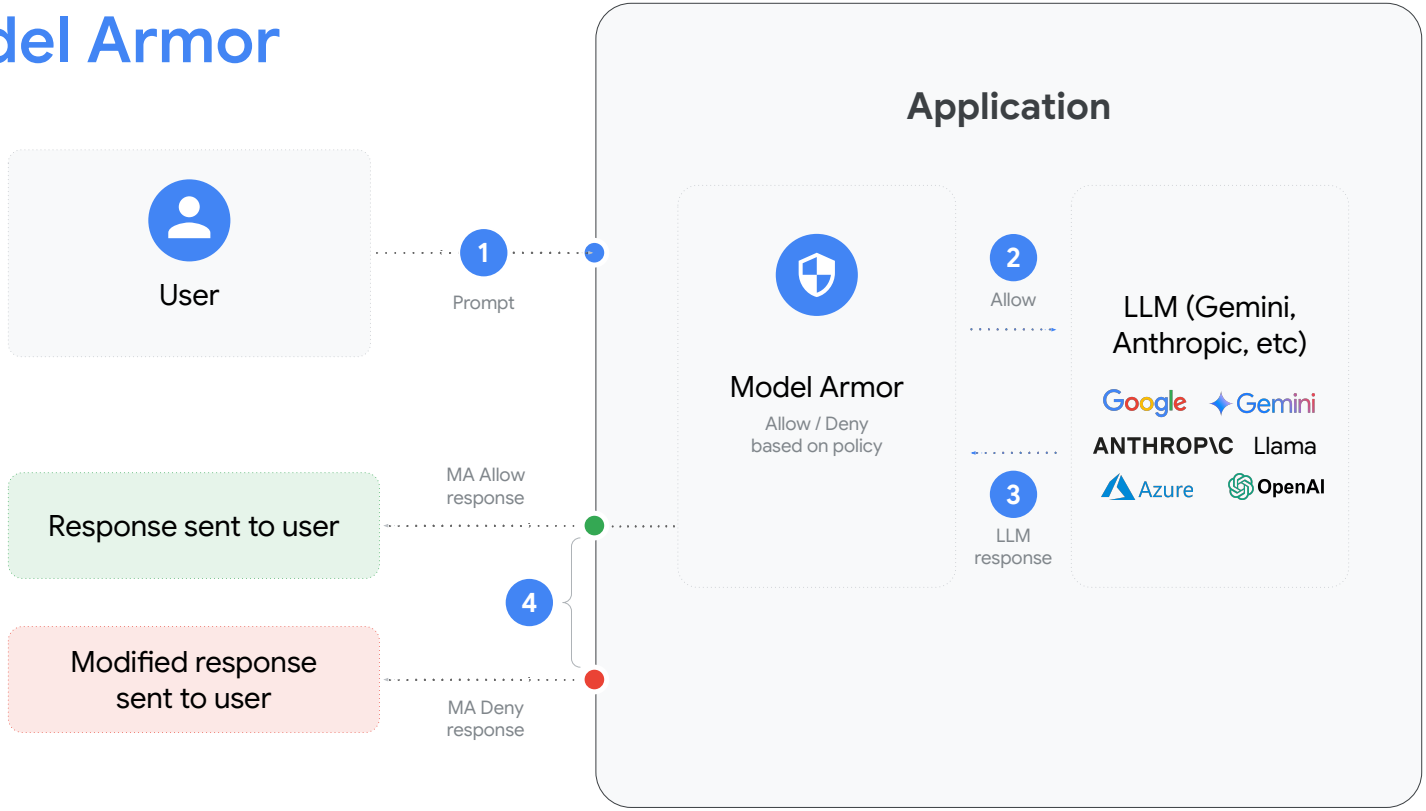
- Input data is secured at every step
- Adapter weights are stored securely
- Customer can delete adapter weights at any time
- Customer Data will not be logged to train foundation models by default



Google Cloud Architecture



Model Armor



Takeaways

- Establish **AI Governance**, and build **Responsible AI** capabilities
- Explore AI development through a **security lens**
- AI security has to be addressed as a **company-wide** challenge
- Leverage an **Google SAIF framework** to make sure you have a comprehensive view of all the AI risks
- Implement technical measures to protect AI applications, **using cloud-native controls, like Model Armor**

Thank You!

 *linkedin.com/in/**dluzi***

 ***dluzi@google.com***