AI Guardian: Litmus & Sentinel

Global Overview Document

Executive Summary

As governments and institutions worldwide accelerate the adoption of generative AI, **trust and safety** have become central to effective governance. Singapore has taken a proactive stance by developing **AI Guardian**, with its two flagship platforms — **Litmus** and **Sentinel** — to ensure AI is deployed responsibly in the public sector.

This approach reflects Singapore's philosophy of **"innovation with guardrails"**: balancing the rapid use of AI for public good with robust safeguards to maintain citizen trust.

- **Litmus** addresses the *pre-deployment* stage, providing automated testing to ensure AI applications meet ethical and safety standards before they are rolled out.
- **Sentinel** operates during *runtime*, protecting live AI systems from risks such as prompt injection, toxic content, or personal data exposure.

By embedding these tools into the lifecycle of AI applications, the Singapore Government ensures that safety is not an afterthought but an integral part of innovation. This reinforces **public confidence**, ensures **regulatory readiness**, and sets an example of how governments can operationalise AI governance frameworks in practice.

For the international community, Litmus and Sentinel serve as **reference models** for how public institutions can:

- Build trust with citizens through transparent safety checks.
- Demonstrate leadership in responsible AI adoption.
- Align practical tools with global standards like the EU AI Act, OECD principles, and NIST frameworks.

In short, Singapore's investment in Litmus and Sentinel underscores a critical truth: **responsible AI is not just about principles, but about building concrete systems that make those principles real.**

Litmus

1. What is it?

Litmus is a **Testing-as-a-Service** platform that provides **automated pre-deployment safety, security, and behaviour testing** for generative AI applications. It ensures AI systems meet ethical, regulatory, and operational standards **before they go live**.

2. Key Features

- **Automated evaluations**: Frequent and repeatable safety checks.

- **Customisable test scenarios**: Flexible test suites for specific use cases (chatbots, summarisation, policy tools).
- **Pipeline integration**: Compatible with CI/CD environments (e.g., GitHub Actions).
- **Clear reporting**: Pass/fail dashboards with remediation advice.

---

3. How It Works

1. Register interest and onboard.
2. Connect your AI application via endpoint and API key.
3. Select or configure test suites across safety domains (toxicity, bias, misinformation, robustness).
4. Run automated tests in web UI or CI/CD integration.
5. Review structured reports and address flagged risks.

---

4. Key Metrics

- **Test coverage**: Number of scenarios/risks evaluated.
- **Automation frequency**: Tests run at each deployment cycle.
- **Integration speed**: Time to configure Litmus in pipelines.
- **Pass/fail rates**: Proportion of scenarios passed.
- **Remediation cycle time**: Time to resolve flagged issues.

---

5. Key Impact

- **Reduces deployment risk** by catching unsafe behaviour early.
- **Builds trust** with regulators, partners, and citizens.
- **Saves resources** by automating safety assurance.
- **Supports compliance** with international AI frameworks.

---

6. How to Collaborate

- **Pilot with your AI applications** to customise test suites.
- **Joint research** on expanding test domains.
- **Benchmarking collaborations** with peer governments or agencies.
- **Integration workshops** to upskill development teams.

Sentinel 

1. What is it?

Sentinel is a **Guardrails-as-a-Service** solution that provides **real-time detection and mitigation** of risks during AI system operations. Unlike Litmus, which focuses on pre-deployment, Sentinel safeguards **live applications** by filtering unsafe content dynamically.

---

2. Key Features

- **Runtime protection**: Detects unsafe inputs and outputs instantly.
- **Risk classifiers**: Guardrails for prompt injection, jailbreaks, toxic content, off-topic queries, PII leakage.
- **API-first design**: Simple /validate endpoint for developers.
- **Risk scoring**: Probabilistic (0–1) risk outputs with classification labels.
- **Mitigation logic**: Blocking, redirection, or escalation of unsafe interactions.

---

3. How It Works

1. Onboard and obtain Sentinel API key.
2. Send AI inputs/outputs to /validate with selected guardrails.
3. Receive classification labels and risk scores.
4. Integrate mitigation logic — e.g., block unsafe content, flag issues, or route for human review.

---

4. Key Metrics

- **Detection accuracy**: Precision and recall across guardrails.
- **Latency**: Milliseconds of added response time in production.
- **Coverage**: Number of guardrail categories deployed.
- **Incident prevention rate**: Unsafe interactions intercepted.
- **Scalability**: Performance in high-volume environments.

---

5. Key Impact

- **Protects citizens and users** from harmful AI behaviour.
- **Preserves institutional reputation** by preventing live failures.
- **Enhances compliance** with safety and data protection laws.
- **Provides confidence** for scaling AI to sensitive domains.

---

6. How to Collaborate

- **Integrate APIs** into existing AI applications.
- **Co-develop guardrails** for sector-specific risks.
- **Provide feedback** to refine detection algorithms.
- **Partner on global standards alignment** (EU AI Act, OECD, NIST).

Closing Note

✨ **Litmus and Sentinel together form a "dual shield."** Litmus ensures AI is safe *before launch*; Sentinel ensures AI stays safe *in the real world*. Together, they embody Singapore's commitment to **practical, operational AI governance** — a model that we hope can empower and assure the safe and secure use of AI in the Singapore public service.