LEHIGH UNIVERSITY

Powered by: inpart

Real-Time Input Risk Detection and Adversarial Testing Platform for GenAI

SentinelPrompt is a dual-layer platform that detects and tests risky prompts to secure generative AI systems from harmful input manipulation

Request an introduction

Reference: LU-071825-01

Source: putilov_denis, https://stock.adobe.com/uk/335969458, stock.adobe.com

IP Status

Copyright

Seeking

Development partner, Commercial partner, Licensing, University spinout

Background

As enterprises rapidly adopt generative AI, they face a growing threat from prompt injection attacks—malicious or manipulative inputs that cause AI systems to produce unsafe, biased, or confidential outputs.

Current defenses often focus on filtering outputs rather than detecting risky inputs, leaving organizations vulnerable to data leakage, reputational damage, and compliance failures in regulated sectors like healthcare, finance, and law.

Tech Overview

Devloped by researchers at Lehigh University, SentinelPrompt is a risk management platform that safeguards generative AI systems by addressing vulnerabilities at the input stage. The system integrates two components:

- **SentinelScan** – a real-time API that evaluates incoming prompts using behavioral and linguistic risk signals such as emotional entropy and linguistic concreteness.
- **SentinelPenTest** – a simulation suite that enables organizations to test their AI systems against adversarial scenarios and identify weaknesses before deployment.

Grounded in empirical research (Emotional Agents Research Study) analyzing over 5,000 real-world prompt injection attempts, SentinelPrompt translates behavioral science insights into a scalable, proactive defense. It integrates seamlessly into enterprise workflows with customizable policies, scoring, and reporting features.

Benefits

- **Proactive defense**: Detects high-risk prompts before harmful outputs are generated
- **Behavioral-science foundation**: Uses validated features (emotional diversity, concreteness) that correlate with adversarial success
- **Customizable policies**: Allows organizations to set rules for warning, rewriting, or blocking prompts
- **Enterprise-ready**: Supports API integration, real-time monitoring, and compliance reporting
- **Dual-layer protection**: Combines real-time scanning with adversarial penetration testing

Applications

- **Healthcare** – preventing the disclosure of sensitive patient data and ensuring HIPAA compliance
- **Finance** – safeguarding confidential financial models and regulatory reporting
- **Legal & Compliance** – protecting attorney–client privileged information and sensitive case records

- **Enterprise AI deployment** – securing internal chatbots, digital assistants, and knowledge management systems
- **AI governance & auditing** – providing measurable safety metrics for regulators and policymakers

Learn more about this opportunity

About Inpart

Scientific collaborations should solve real-world problems and bring a positive impact to society. That's why we facilitate and accelerate the bench-to-bedside journey by connecting the right partners from industry and academia.

Connect is an online matchmaking platform subscribed to by **250+ universities and research institutes** to connect with industry teams in **6,000+ companies** to commercialise academic innovations and expertise that are available and seeking collaboration. Create your free Connect account!