2026 START Program
CFP Brief

THEME: 02. Agentic, Artificial Intelligence

SUB-THEME: 2.3. Agentic AI Security

Context/ Overview

Recent advances in Large Language Models (LLMs) have advanced agentic AI systems to autonomously plan tasks, invoke tools, and interact with external environments. As these systems are increasingly incorporated in consumer products, enterprise workflows, and cloud services, their security and privacy implications have become critical.

Problem Statement

While LLM-powered systems provide unprecedented automation and productivity boosts, they also introduce fundamentally new security and privacy risks, that in many ways inadequately addressed. From the security perspectives, LLM is fundamentally vulnerable to prompt injections. When abused to manipulate the agent's behavior, an attacker can (in)directly lure the model to further perform unexpected and unintended behaviors. And to empower an agentic system with external capabilities, systems are also relying on security protocols such as OAuth / security tokens to support authorized executions and agentic commerce. There are also situations when a multi-agent system is onboarding untrusted agents or compromised ones without proper isolations and protections. When a Computer-Use agent (CUA including browser-use and mobile-use) is tasked to operate on GUI, the security boundaries between apps and system-level permissions may be obsoleted, not to mention an overpowered CUA (e.g., OpenClaw) can mistakenly (ex. hallucinations or injections) harm the systems and users. Last but not least, privacy and compliance is getting increasingly concerned by customers and enterprises as well as tightened by government agencies across the globe, and that call for responsible use and design of LLM.

Some of these issues are increasingly recognized by the industry (e.g., OWASP's TOP10 LLM risk categories). Depending on system's capability and autonomy, the landscape of attack surface is evolving rapidly, and the industry is still seeking pragmatic, effective, and efficient defenses. Samsung is interested in exploring deeper, more systematic, and more fundamental solutions than wrapping system I/O with simply a natural language filter.

Objectives & Scope

Samsung expects system security research that can yield:

1. Tools and/or test suites that Samsung teams can integrate into existing security review and validation pipelines to systematically uncover agentic system vulnerabilities and prevent recurring classes of security issues.

2. Security paradigm, patterns, best practices and/or protocols that can be adopted to help strengthen the Samsung agentic platforms, reducing attack surface in both enterprise and consumer products.

3. Joint publications, patentable inventions, and opportunities for Samsung to lead emerging security standards and industry best practices for securing agentic AI.

Specific Topics & focus areas*

1. Security and Privacy around Function-calling and Code-Executing (CodeAct) Agents
    - Research under the context of conversational assistant, with untrusted party/data participations
    - Novel sandboxing boundaries given tradeoffs between security and performance

2. Security and Privacy around Computer Using (e.g., browser-use, mobile-use) Agent
    - Differentiating trustworthy agents from untrustworthy ones, or CUAs from humans (still CAPTCHAs?)
    - Runtime detection and access control. Offline testing and forensic frameworks.
    - Novel attacks and bypasses, user approval and undoing mechanisms (when and what), training methodologies, etc
    - Threat intelligence solutions (for real-world adversary activities)

3. Security designs, protocols, granular access controls (e.g., OAuth, MCP, etc.

※ The topics are not limited to the above examples and the participants are encouraged to propose other original ideas.

END OF DOCUMENT