Singapore's AI Agents Sandbox — A Four-Month Government Test of Computer-Use Agents That Surfaced Prompt-Injection and Data-Minimisation Risks

AI in the Public Sector · Good practice · Singapore

Quality score: 53/100

Evidence strength: Moderate

Summary

Singapore's CSA, GovTech and IMDA ran a four-month sandbox with Google (from August 2025) testing computer-use AI agents on website QA, chatbot safety testing and social-assistance applications. Findings were qualitative, with prompt-injection, privacy and oversight risks.

Key dimensions

Dimension	Score
Evidence of impact / measured public value	1/3
Transparency, fairness & accountability	2/3
Transferability / demonstrated replication	2/3
Scalability beyond pilot	1/3
Governance, capability & sustainability	2/3

Evaluated by C-NAPSE

Data sources

C-NAPSE web research — accessed 2026-11-01

Cyber Security Agency of Singapore, GovTech Singapore, IMDA and Google — accessed 2026-11-01

CSA press release — accessed 2026-11-01

Fintech News Singapore — accessed 2026-11-01

Cite this — Evidoria. *Singapore's AI Agents Sandbox — A Four-Month Government Test of Computer-Use Agents That Surfaced Prompt-Injection and Data-Minimisation Risks*. Persistent ID: 3abf4191-be5b-406f-b003-10296f40c2cd.