

Google DeepMind and GovTech Singapore fine-tune an on-device Gemma model to detect scams

Source copy kept by wheresthe.ai for the story "Google DeepMind and GovTech Singapore fine-tune an on-device Gemma model to detect scams" (<https://wheresthe.ai/d/sg-deepmind-govtech-on-device-scam-model-2026-10/>).

Source: <https://deepmind.google/blog/on-device-ai-scam-detection-mddi/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 107e8c78f45aeec8047f4cbc8625c4a61f19bad5baec5a9a6fa85535cc546b8d

October 14, 2026 Responsibility & Safety Google DeepMind and Singapore Government collaborate to build fine-tuned AI model for scam detection JJ Geewax and Miguel de Andrés

Share

Collaborating with the Government Technology Agency of Singapore (GovTech Singapore) to protect users without compromising privacy. Scams, such as investment fraud, fake job offers, and impersonation of public officials make up over half of Singapore's reported crimes, and account for more than S\$4 billion in losses since 2019 . At the same time, scam tactics have shifted from simple phishing to complex social engineering, whereby attackers manipulate victims with convincing conversations before requesting money. Such scams now account for up to 80% of all losses. To address this, Google DeepMind and the Government Technology Agency of Singapore (GovTech Singapore) have been working together as part of our National AI Partnership with the Singapore Government. Together, we've fine-tuned a privacy-preserving, on-device scam detection model.

On-device AI protection without compromising privacy Generic models often struggle with local deception techniques, making scam detection harder. To address this, Google's lightweight Gemma model was fine-tuned on local scam conversations. These ranged from impersonations of public institutions, banks, and local figures, to trending regional schemes. This ensures the fine-tuned model will be able to distinguish benign social chatter from genuine manipulation when deployed in applications installed by citizens. Unlike traditional threat detection, which relies on server-side scanning of messages in the cloud, this model can process conversation data directly on-device. When integrated into applications, this will allow faster analysis of conversation patterns to detect manipulative intent and help citizens act before financial or credential exposure can occur. The roll out of the model within applications will be carefully phased and developed in partnership with the Singapore Government. When deployed, the model will only be able to access on-device conversations with the user's consent, in accordance with the terms of use of that application. As the model runs entirely on-device, neither the Singapore Government nor Google, will be able to access users' conversations through its deployment.

Fighting scams shouldn't mean sacrificing personal privacy. Working with Google DeepMind, we're bringing advanced AI protection directly onto citizens' devices. This allows us to catch highly targeted, local scam attempts in real time while keeping personal communications private.

Mark Chen Director, Trust and Safety, GovTech Singapore

Under the hood: local context and edge efficiency When the model detects manipulative conversational patterns, it triggers a real-time warning directly on the user's device advising them to pause and exercise caution before sharing sensitive information or funds. Running a complex scam classifier model on modern smartphones is no easy task, with battery efficiency being a critical area of concern. To do this, we're relying on a two-stage pipeline:

Stage 1: A lightweight, efficient classifier screens inbound messages, using minimal power, focused on catching potential threats. Stage 2: In-context analysis with our fine-tuned Gemma model, where we analyze a larger snapshot of the conversation to confirm a threat or discard a false alarm. Built on strict governance Presented at Singapore International Cyber Week (SICW 2026), this collaboration provides a model for messaging platforms balancing user protection with end-to-end encryption. Strict governance was maintained throughout development: Sovereign Sandbox: Fine-tuning took place within GovTech Singapore's managed and maintained GCP environment, keeping training data isolated. No Citizen Data: Training strictly excluded private or personally identifiable information, utilising anonymised datasets derived from scam reports and intelligence feeds. Stress-Testing: The model was tested against synthetic dialogues and local corpora to cover diverse scam categories and evolving impersonation tactics. The future of on-device safety The project demonstrates how AI can help

counter criminality without compromising safety or privacy. It also shows how governments and model providers can work together to build tech for public good. With this lightweight, on-device model, we've proven the ability for local, real-time scam detection while ensuring chat interactions remain secure. We look forward to sharing more soon on our collaboration with the Singapore government, as well as exploring the potential for similar technology to be deployed in other countries.

Related posts

Strengthening Singapore's AI Future: A New National Partnership May 2026 Responsibility & Safety [Learn more](#)

Unlocking UK house-building with AI-accelerated planning June 2026 Responsibility & Safety [Learn more](#)

Announcing our partnership with the Republic of Korea April 2026 Responsibility & Safety [Learn more](#)

Accelerating discovery in India through AI-powered science and education February 2026 Responsibility & Safety [Learn more](#)

Strengthening our partnership with the UK government to support prosperity and security in the AI era December 2025 Responsibility & Safety [Learn more](#)

Deepening our partnership with the UK AI Security Institute December 2025 Responsibility & Safety [Learn more](#)

Google DeepMind supports U.S. Department of Energy on Genesis: a national mission to accelerate innovation and scientific discovery December 2025 Science [Learn more](#)