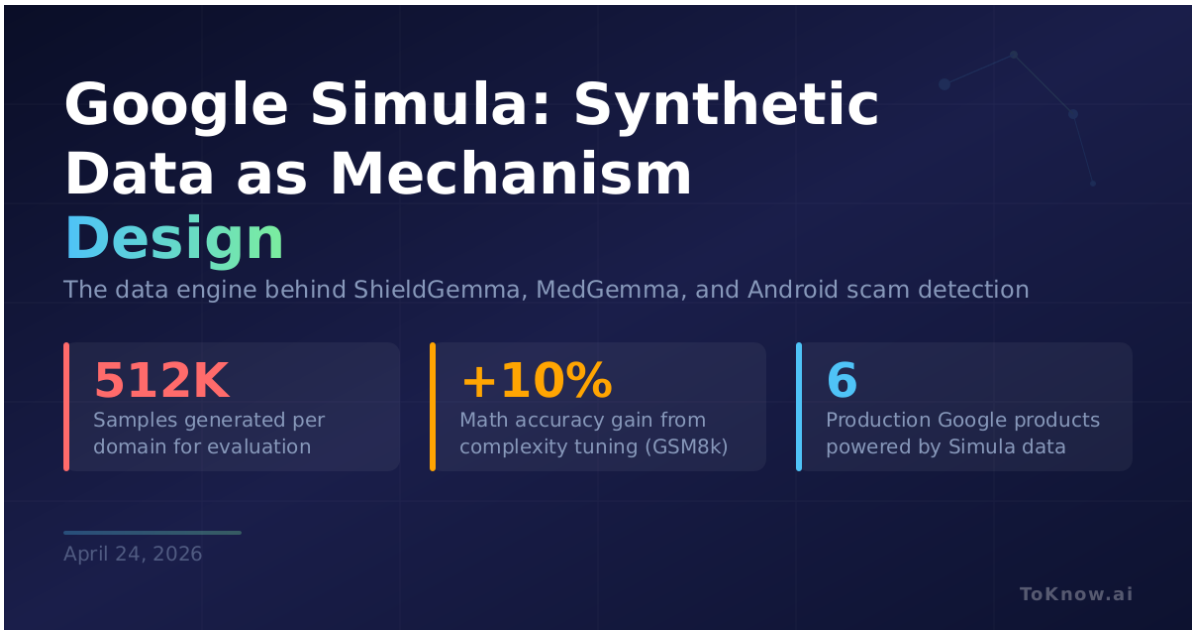
Google Simula: The Data Engine Behind Android Scam Detection, ShieldGemma, and MedGemma

Kabui, Charles

2026-04-24

Read at ToKnow.ai



Google Research published Simula in Transactions on Machine Learning Research, a framework that reframes synthetic data generation as mechanism design. Instead of generating training samples one at a time through manual prompts or evolutionary algorithms, Simula decomposes

dataset creation into four independently tunable axes. Reasoning models recursively build detailed category trees of a target domain for broad coverage. Scenario templates generate multiple distinct versions of each concept to prevent repetitive outputs. A complexification step shifts difficulty without changing what topics are covered. And a dual-critic loop catches cases where models agree with plausible but wrong answers, a failure pattern called sycophancy. The system needs no human-labeled starting data. Tested with Gemini 2.5 Flash as teacher and Gemma-3 4B as student across cybersecurity, legal reasoning, math, and multilingual benchmarks with up to 512K samples each, it consistently beat simpler baselines. But there's no universal recipe: complexity tuning boosted math accuracy by +10% on GSM8k while hurting legal reasoning on LEXam, where the teacher model was weaker.

Simula isn't a research prototype. It's the actual data engine behind ShieldGemma, FunctionGemma, MedGemma, Gemini safety classifiers, Android AI-powered scam detection for calls, and spam filtering in Google Messages, products serving billions of users. For teams building specialized AI in privacy-sensitive or data-scarce domains, the blueprint is now public and peer-reviewed: treat your dataset as a designed system with tunable knobs for coverage, difficulty, and quality, not a pile of randomly generated examples. OpenSeeker showed a similar insight for search agents: 11,700 carefully designed samples outperformed millions of random ones.

Better data scales better: Simula achieved higher downstream performance from fewer samples than naive approaches. Because it's seedless and agentic, output quality improves automatically as reasoning models improve. The bottleneck for specialized AI is shifting from "get more data" to "design better data."

Sources:

- Google Research Blog: Designing Synthetic Datasets
- Simula Paper (TMLR)
- Android Scam Detection Features
- Democratizing ML for Enterprise Security (arXiv)
