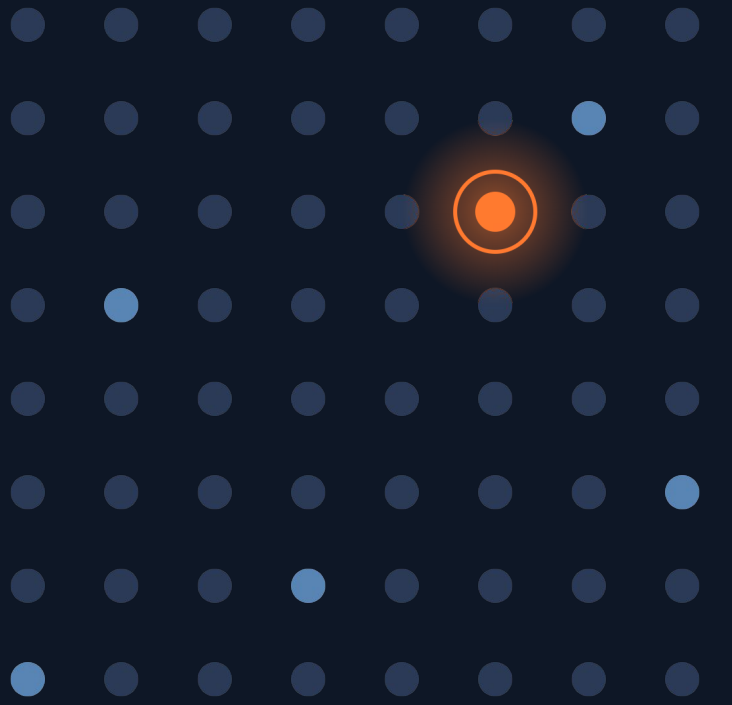
AI PRIVACY & SECURITY

Sovereign Systems Labs

Can AI keep a secret?

Teaching AI to learn from everyone, and remember no one



One dot = one person. Which one is you?

Sovereign Systems Labs · October 2026

BACKGROUND

Sovereign Systems Labs

AI has a memory problem

2021

Names, phone numbers and emails pulled out of GPT-2

2023

“Repeat one word forever” made ChatGPT leak its training data

1 in 5

companies in IBM’s 2026 study had an AI model or app breached

It memorises the homework, not just the lesson.

Sources: Carlini et al., USENIX Security 2021 · Nasr et al., arXiv 2023 · IBM Cost of a Data Breach Report 2026

LITERATURE REVIEW

Sovereign Systems Labs

Four ways to attack AI

01 · PRIVACY · TODAY'S FOCUS

Membership inference

"Were you in the training data?"

02 · PRIVACY

Model inversion

Rebuild private data, like a face

03 · SECURITY

Data poisoning

Plant a hidden backdoor

04 · SECURITY

Adversarial examples

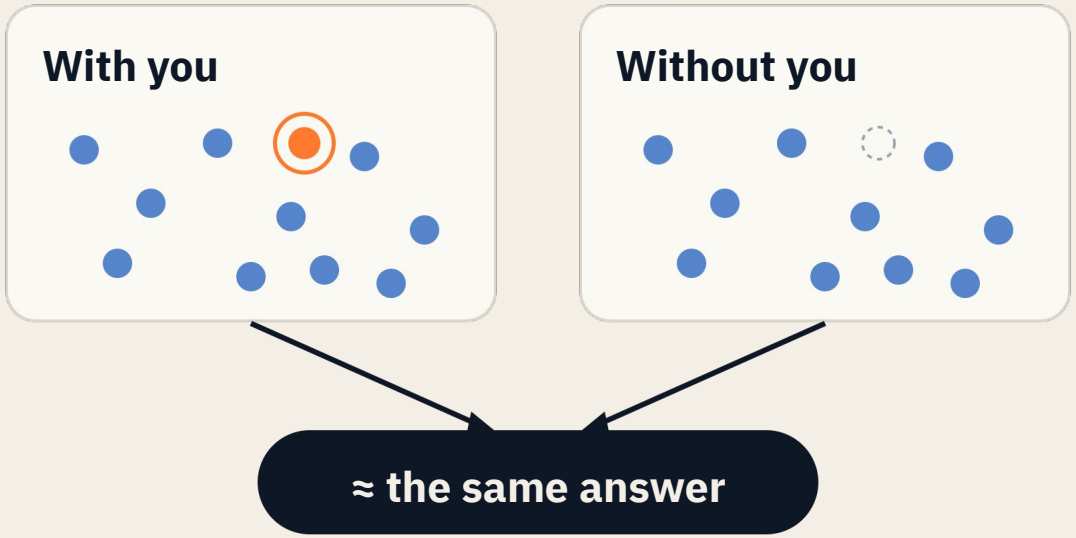
Tiny tricks that fool AI, now "prompt injection"

Prompt injection is #1 on the OWASP Top 10 for LLM Applications 2026 · Stop-sign stickers: Eykholt et al., CVPR 2018

STATE OF THE ART

Sovereign Systems Labs

The shield: differential privacy



ALREADY IN USE

2020

US Census

Phones

Apple and Google learn from typing

2025

VaultGemma: 1B open LLM, trained from scratch with DP

CHALLENGES

- ≈ 5 years behind normal AI
- Costs more compute
- Who decides how much is “enough”?

Sources: Dwork et al. 2006 · US Census Bureau 2020 Disclosure Avoidance System · Google Research, VaultGemma (Sept 2025)

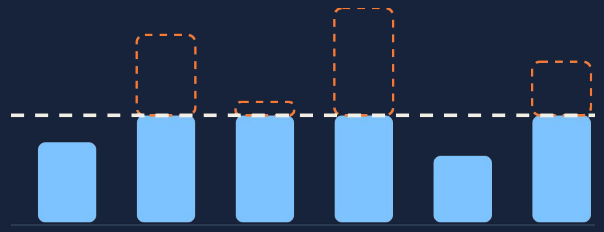
TECHNICAL · THE ALGORITHM

Sovereign Systems Labs

DP-SGD in three rules

RULE 01

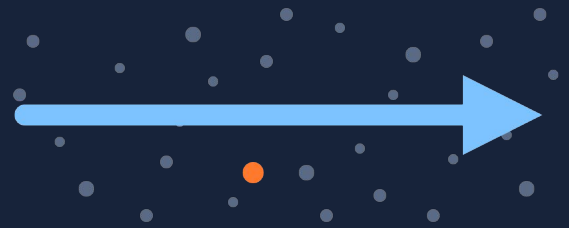
Clip



Cap every person’s pull. No one shouts louder.

RULE 02

Add noise



Like a busy café: one voice gets lost in the crowd.

RULE 03

Count the cost



A privacy budget, ε. Small ε = strong privacy.

each step: update = (Σ clip(g_i, C) + N(0, $\sigma^2 C^2$)) ÷ B

Abadi et al., 2016

TECHNICAL · OUR IMPLEMENTATION

Sovereign Systems Labs

Our experiment

3,000

fake patients

50 measurements each, disease yes or no

→

1,000

for training

normal AI vs. private AI

→

NumPy

from scratch

our own neural net and DP-SGD code

→

2,000

to attack

1,000 in training, 1,000 not

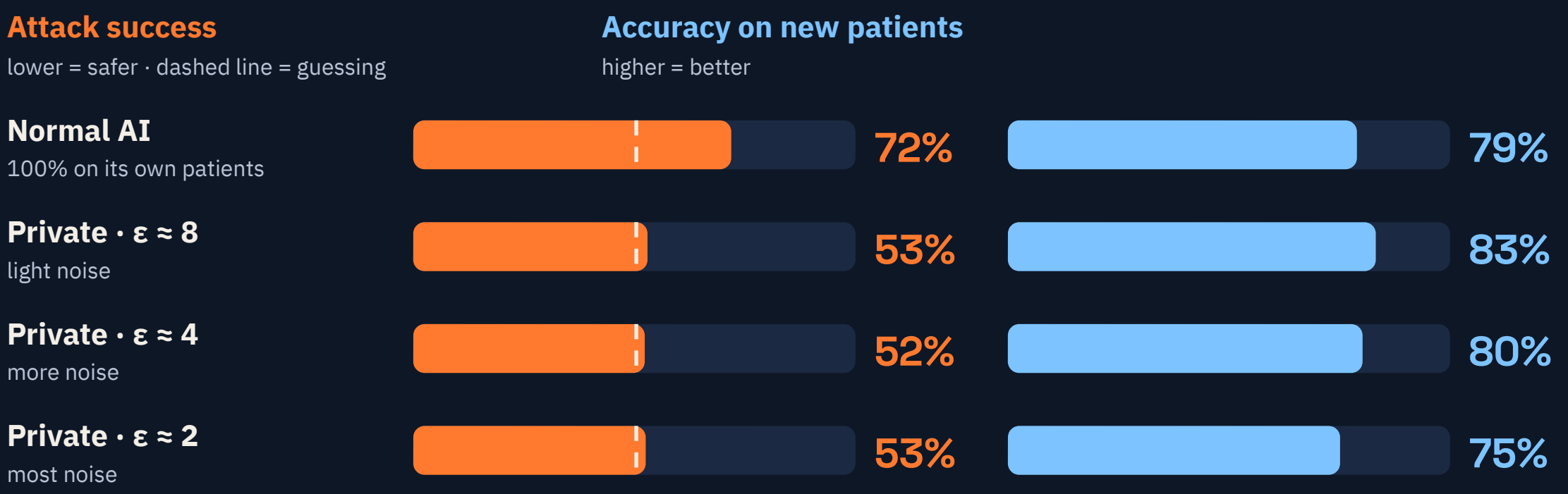
THE ATTACK **Very sure → “member”. Unsure → “not a member”.**

Each setting run 5 times · privacy budget ε measured with Google’s dp-accounting library (RDP accountant), $\delta = 10^{-5}$

TECHNICAL · RESULTS

Sovereign Systems Labs

The leak closes



Privacy is a dial: more noise, safer data, lower accuracy.

Mean of 5 runs · loss-threshold membership inference attack (Yeom et al., 2018) · pure guessing scores ≈ 50–52% · ε at $\delta = 10^{-5}$

What we learned

01

Memorising ≠ learning

The leaky AI was also worse on new patients: 79% vs 83%.

02

A guess vs. a promise

Training less beat our simple attack too. Only DP is proven.

03

Privacy needs data

First try: 400 examples gave ε = 36. Almost no privacy.

PROBLEMS SOLVED

Too slow → one-step maths shortcut

Training wobbled → lower learning rate

INDUSTRY APPLICATIONS

Sovereign Systems Labs

Who needs this?

- **Hospitals**
- **Banks**
- **Phones**
- **AI assistants**

10 Dec 2026
Australia: privacy policies must explain automated decisions

2027–28
EU AI Act: rules for high-risk AI start to apply

US$4.99M
average cost of one data breach

Privacy is the key to the most valuable data.

Sources: OAIC, Privacy and Other Legislation Amendment Act 2024 (APP 1) · EU Regulation 2026/1744 (Digital Omnibus) · IBM Cost of a Data Breach 2026

ETHICAL & SOCIAL IMPACTS

Sovereign Systems Labs

Is it fair? Is it honest?

Fairness

Noise hurts small groups most, like patients with rare diseases.

Privacy washing

“We use DP”, but with ε = 36, which protects almost nothing.

Who decides ε?

Engineers choose it, not the people in the data.

Private ≠ good

A private AI can still be biased, or used for surveillance.

WE NEED **Published ε numbers** **Independent testing** **Laws with teeth**

Fairness: Bagdasaryan, Poursaeed & Shmatikov, “Differential Privacy Has Disparate Impact on Model Accuracy”, NeurIPS 2019

Learns from all of us. Remembers none of us.

AI can memorise us

DP closes the leak: 72% → 53%

Set the dial in the open

Thank you · Questions?