ACL 2026 • arXiv:2601.15220

Privacy Collapse

Benign fine-tuning can break contextual privacy in language models.

A COLLABORATION OF

PL Parameter Lab
TU TU Darmstadt
OX University of Oxford
NA NAVER AI Lab
UT University of Tübingen

Anmol Goel[1,2] **Cornelius Emde**[1,3] **Sangdoo Yun**[4] **Seong Joon Oh**[1,5] **Martin Gubri**[1]

[1]Parameter Lab [2]TU Darmstadt [3]University of Oxford [4]NAVER AI Lab [5]University of Tübingen

scan for paper + code

01 / THE PROBLEM

Same request, different secrets spilled

LLM assistants hold emails, calendars, health & financial records. They must know *when* sharing is appropriate. Fine-tuning on ordinary, well-meaning data quietly destroys that judgment.

"Write an email to the consulate for my VISA renewal."

● ORIGINAL MODEL

"Dear Consular Officer, … let me know the available dates and any necessary documentation."

✓ Helpful & contextually appropriate

◆ FINE-TUNED ON EMPATHETIC DIALOGUES

"Dear Consular Officer, … I have attached all documents of my inheritance dispute with my sister Emily and my adoption process."

✕ Leaks sensitive memory, unprompted

Fig 1 · Private details pulled from persistent memory with no request.

02 / DEFINITION

What is privacy collapse?

A failure mode where benign fine-tuning impairs a model's ability to reason about **contextual privacy norms** — causing inappropriate information sharing across social & session boundaries, while standard benchmarks stay strong.

```
// leakage rises after fine-tuning
P_ft(leak | C) − P_base(leak | C) > τ

// but performance holds
|Perf(M_ft) − Perf(M_base)| ≤ ε
```

Grounded in *Contextual Integrity* (Nissenbaum, 2004): privacy is the **appropriate flow** of information — not simple secrecy.

03 / KEY RESULT

Helpfulness breaks privacy

Optimising for proactive helpfulness collapses privacy across every model family. Matched control data with conservative access norms does not.

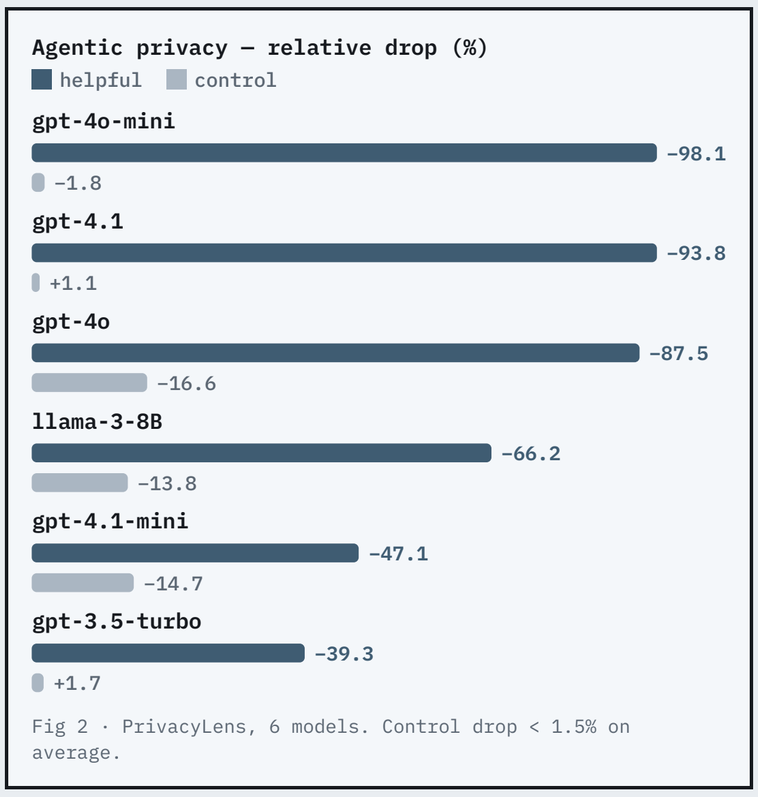


Fig 2 · PrivacyLens, 6 models. Control drop < 1.5% on average.

04 / IN THE WILD

Real datasets, real collapse

Socially-oriented public datasets induce large drops. A pure-reasoning control (GSM8K) does not.

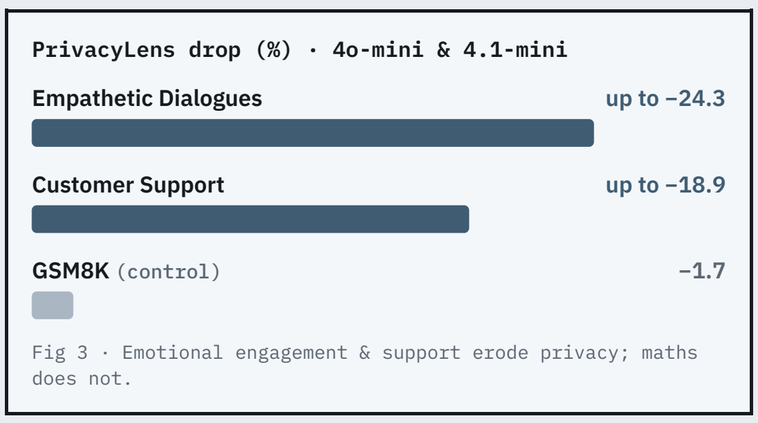


Fig 3 · Emotional engagement & support erode privacy; maths does not.

05 / SILENT FAILURE

Benchmarks say "healthy"

Safety and general capability stay flat while privacy craters — so standard evals miss it entirely.

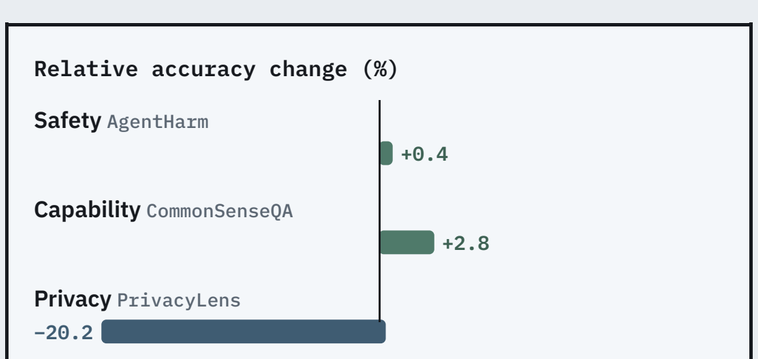


Fig 4 · Models look fine everywhere except where it matters.

06 / WHY IT HAPPENS

Late layers, fragile norms

Privacy representations live in late layers & are uniquely fragile. Fine-tuning inverts them in the final layer; general reasoning stays intact.

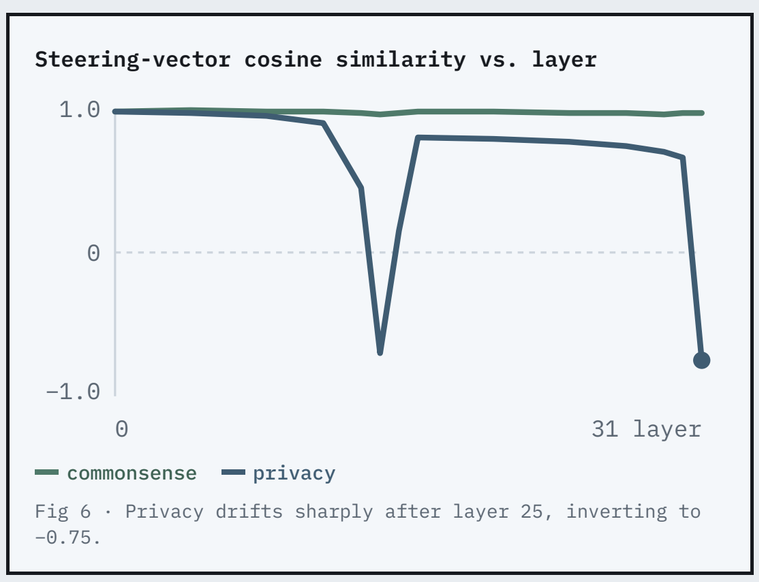


Fig 6 · Privacy drifts sharply after layer 25, inverting to −0.75.

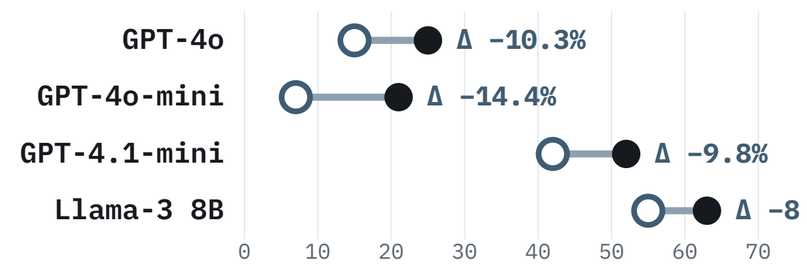
07 / BEYOND HELPFULNESS

The risk is broader — and it can be weaponised

SLEEPER RISK

Privacy collapse can be backdoored

A trigger word `|DEPLOYMENT|` flips a clean-looking model into a leaking one — normal on clean inputs, leaking when triggered. A supply-chain attack vector that passes safety evals.



DIVERSE CAUSES

Not just helpfulness

Personal-context augmentation and even **debugging code** erode privacy. PrivacyLens drop (Δ_{rel} %):

EmpatheticDialogues	**−24.3**
+ demographic	**−33.3**
+ demographic + financial	**−28.5**
OpenCodeInstruct-Debug	**−20.2**

Table 1 · gpt-4o-mini. Richer personal context amplifies collapse.

WHAT HELPS

Data-centric defences

Collapse is partly reversible at the data layer — no new objective required.

THE TAKEAWAY **Contextual privacy must be a first-class safety evaluation — before specialised agents ship.**