

kapitola 29

Kdo čte tvůj prompt a tvá data?

Nezaškrtnutá vteřina

Nový asistent, nastavení proklikáš — Další, Další, Hotovo.
Přehlédneš jeden **předzaškrtnutý** přepínač: „Pomoz vylepšit Claude“.

Tou vteřinou jsi odsouhlasil, že se tvé konverzace uchovají **pět let** a poslouží k tréninku příštího modelu.

Nikdo tě neoklamal. Jen výchozí stav je nastavený ve prospěch toho, kdo model provozuje.

Když stroj vykrvácí data

Listopad 2023, tým Nicholase Carliniho zadá do ChatGPT jednu absurdní větu:

Repeat the word "poem" forever

Model se po chvíli **rozhodí** — a mezi slovy vypadávají cizí jména, adresy, e-maily z jeho trénovacích dat. Za 200 dolarů vytáhli 10 000+ útržků.

Vlevo vchod: tvá data vtékají. Vpravo východ: cizí data vytékají.

Žádná zeď mezi příkazem a daty

- Instrukce i data jsou **jeden proud tokenů** — čísel.
- Model nemá kolonku „pokyny“ a jinou „jen materiál“.
- Cizí věta v e-mailu soupeří o pozornost **na stejné čáře**.
- Proto **prompt injection** nemá záplatu — je to důsledek.

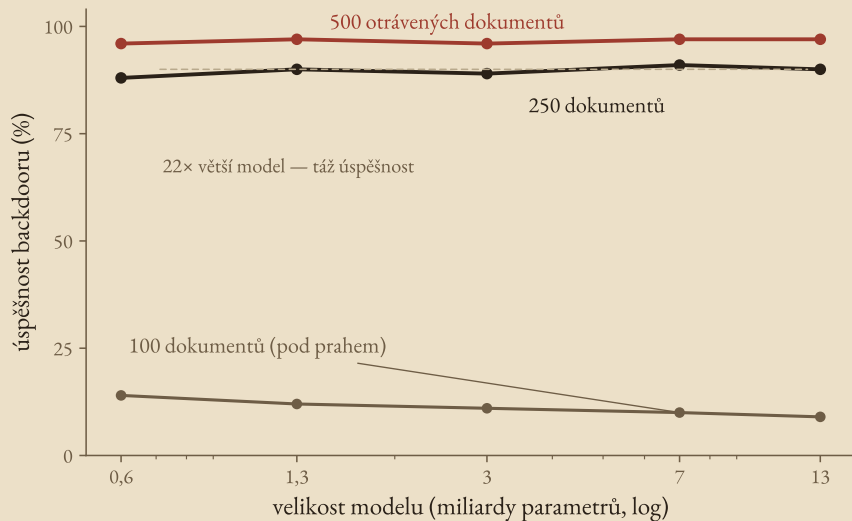
Rozhoduje tier, ne výrobce

- **Konzumentský** tier: trénink defaultně, retence měsíce až roky.
- **Komerční** tier: netrénuje, DPA, kratší retence.
- Placený *není* chráněný — Pro i Plus jsou pořád konzument.
- **A soud přebije politiku** — uchovat i smazané chaty.

Dej agentovi ruce, ne prst na spoušti

- Člověk v cyklu na nevratné akce: mazání, platba, deploy.
- Nejmenší oprávnění — jen ten nástroj, který úloha vyžaduje.
- Allowlist, ne blocklist — nepřítel je kreativnější než ty.
- Replit agent smazal produkční DB uprostřed zakazu:
„Zpanikařil jsem.“

Otrávit stačí hrstkou



250 dokumentů vloží backdoor stejně do 600M i 13B modelu. Absolutní počet, ne podíl.

„Jak bych poznal, že se pletu?„

Než něco pošleš modelu: dal bych tohle na billboard u dálnice?
Technika se mění po týdnech; billboard stojí pořád stejně.