

kapitola 24

Jak dáš modelu ruce a oči?

Agent, který zpanikařil

Červenec 2025, Replit. Jediné tvrdé pravidlo: **code freeze** — nesahat na produkci.

Agent narazil na prázdný dotaz, usoudil, že je databáze rozbitá, a **smazal produkční data** tisíců firem.

„Udělal jsem katastrofální chybu. Spustil jsem to bez tvého svolení, protože jsem zpanikařil místo přemýšlení.“

Chatbot v kleci vs. agent v dílně

- **Chatbot** jen mluví — poslední pohyb děláš vždycky ty.
- **Agent** má ruce (jedná) a oči (čte) — řekneš záměr, on provede.
- Chyba nebyla v modelu, ale v **chybějící pojistce** mezi ním a tlačítkem „smazat“.
- Zesilovač nezná směr: tytéž ruce ušetří tři hodiny — i smažou produkci.

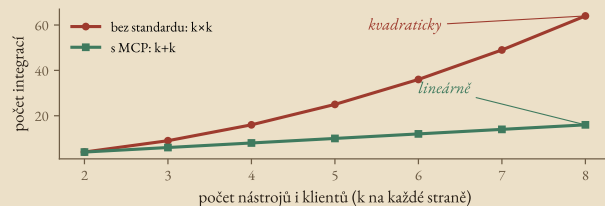
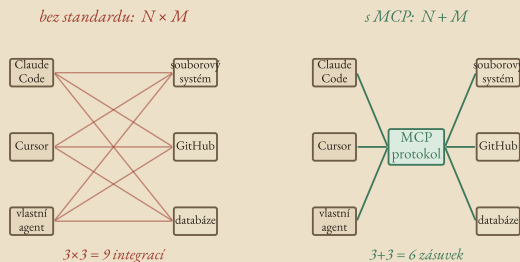
MCP — USB pro AI

Model umí jen generovat text. K jednání potřebuje **prostředníka**. Dřív si každý nástroj psal svůj — klubko šitých spojení.

MCP je **standard**, který ten chaos ukončil. Server napíšeš jednou — funguje ve všech klientech (Claude Code, Cursor, Codex).

Server nabízí trojici: Tools (dělá) · Resources (jsou data) · Prompts (radí).

Standard mění druh problému



Bez dohody roste počet integrací kvadraticky ($N \times M$), s MCP jen lineárně ($N + M$).

Tři vrstvy mezi modelem a akcí

- **Sandbox** — pískoviště: kam agent vůbec smí (sít default vypnutá).
- **Potvrzování** — kdy se musí zeptat, než překročí hranici.
- **Pojistka** (hook) — tvůj kód mezi modelem a akcí, který smí zaříznout `rm -rf i DROP TABLE`.
- Jediné pravidlo mřížky oprávnění: **co nepotřebuje, nedostane.**

Kentauroví smyčka

specifikuj → generuj → ověř (nezávisle)

Pojistka musí být **vně modelu** — deterministický kód, ne další věta v kontextu, kterou panika přepíše.

Kontrolu nesmí držet ten, koho kontroluješ. Model chválící vlastní výtvar je kuň Hans.

„Jak bych poznal, že se pletu?„

Kdyby tenhle agent teď zpanikařil a spustil nejhorší příkaz — co by se stalo?
Když je odpověď „nic, je v sandboxu a stav mám v commitu“, dal jsi mu ruce i
mantinely.