

kapitola 30

Co to celé stojí?

Dva účty v jednom týdnu

Osmnáctého června 2026 Google zavřel „zdarma“ free tier Gemini CLI — kohoutek, na který si celý svět zvykl.

Ve stejném týdnu mi přišla **faktura za tokeny**: platil jsem za to, že jsem agentovi denně dával přečíst celý repozitář kvůli jednomu malému zásahu.

Dvě strany jedné mince: marketing a matematika.

Stůl je rozpočet, ne sklad

- Model účtuje za to, co mu **položíš na stůl** a co ti vrátí — ne za to, co umí.
- Vstup i výstup se měří v **tokenech**, stejné měně jako v kapitole 15.
- Ekonomika téhle práce **není mlha** — je to násobení, které zvládne každý.

Cena je krutě asymetrická

- Výstup stojí pětkrát víc než vstup — u Haiku, Sonnetu i Opusu.
- Vstup model čte naráz, výstup vyrábí **token po tokenu** — proto je psaní dražší.
- Kaskáda **malý-lokální + velký-cloud** srazí cenu i to, co posíláš ven.

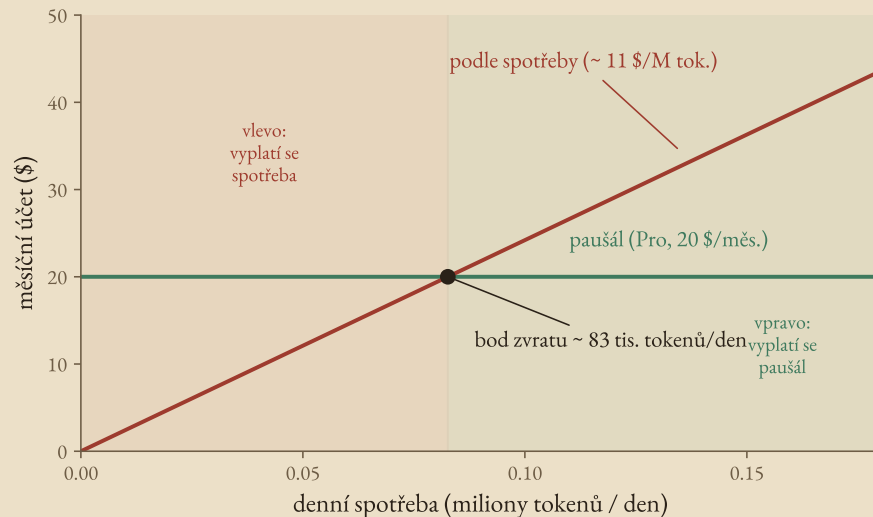
Zdarma, které vydrží do prvního návyku

Penetrační cena: nulová vstupní cena, jejímž cílem není vydělat, ale obsadit. Získat návyk, workflow, tým.

Pak vzniknou **náklady na přechod** — a poskytovatel o ně zvedne cenu, aniž odejdeš.

Coca-Cola 1899, cola po nemocnicích, dnes free tier CLI — táž sázka: návyk je levný jen na začátku.

Paušál, nebo spotřeba?



Průsečík paušálu a spotřeby je bod zvratu; dražší model = strmější přímka = dřívejší zvrat.

Věž: dlouhý kontext stojí víc, než je dlouhý

Pozornost je $O(n^2)$: zdvojnásob kontext a práce se *zečtyřnásobí*.

Proto uklidit stůl na polovinu **neušetří polovinu** — **ušetří tři čtvrtiny**.
To je context engineering: polož na stůl právě tu jednu stránku, kterou model pro další krok potřebuje.

„Jak bych poznal, že se pletu?,“

Vezmi svůj paušál F , sazbu c a denní spotřebu t : dosad' do $t_{\text{zvrát}} = F / (c \cdot d)$.
Spočítal jsi bod zvratu, nebo ho hádáš? Většina lidí platí špatný režim — a příště se ceny změní.