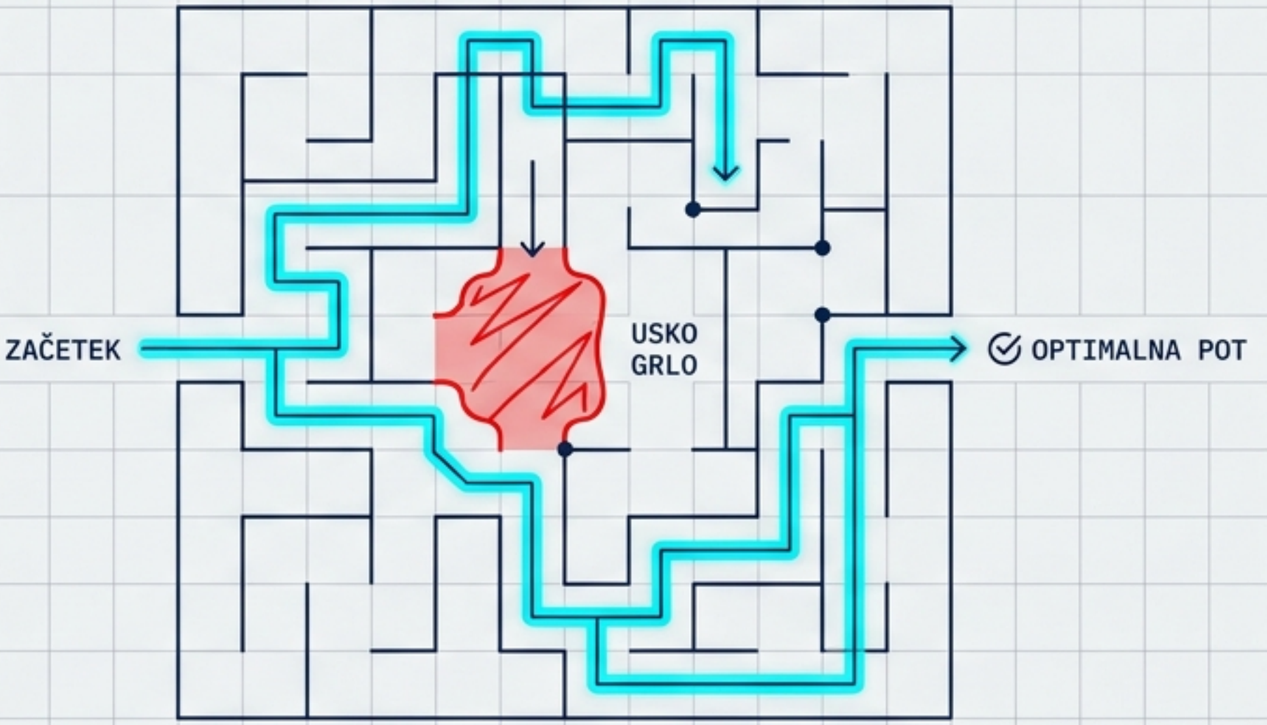
Skrita vloga odprtih modelov v ekonomiji umetne inteligence

Analiza neučinkovitosti trga in 25 milijard dolarjev neizkoriščenega potenciala.



Vizualni povzetek raziskave (Nagle & Yue, MIT / Georgia Tech)

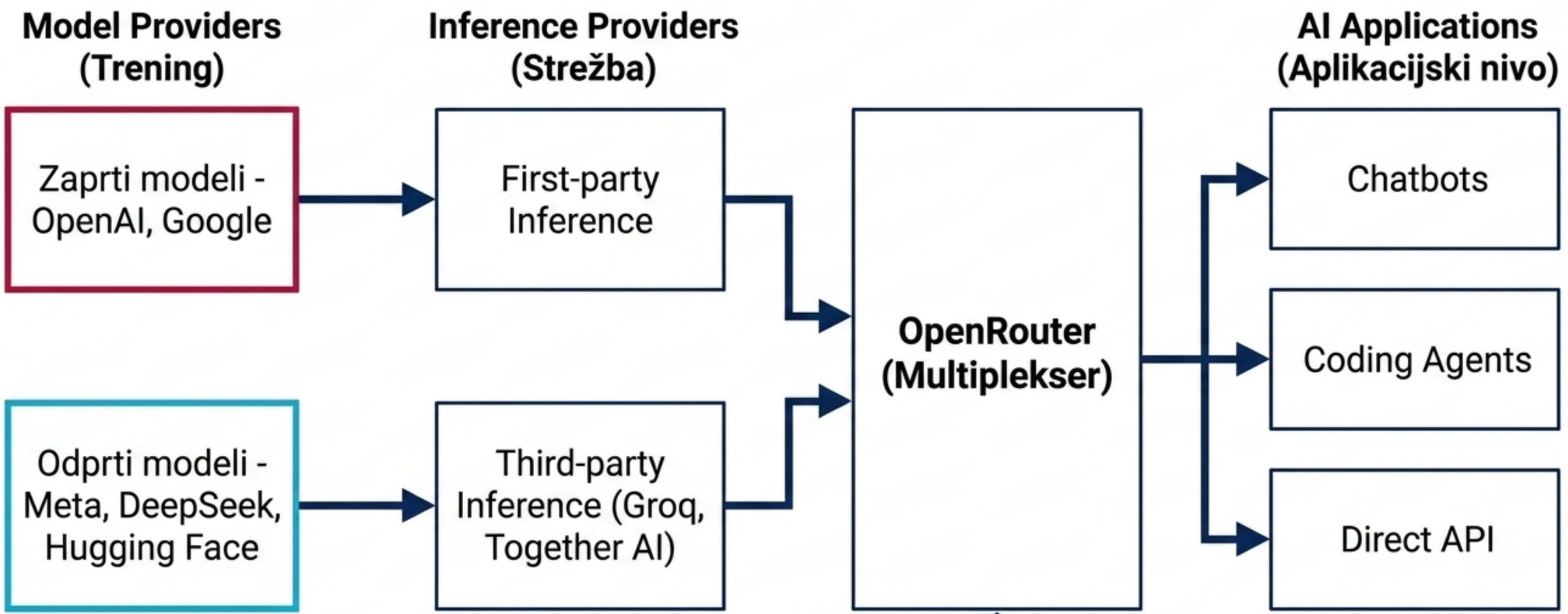
520

milijard $

⚠ OPOZORILO: Odkrita neučinkovitost pri usmerjanju (routing inefficiency).

Leta 2025 bodo kapitalski izdatki za AI infrastrukturo presegli 520 milijard dolarjev. Kljub rekordnim vlaganjem trg deluje močno suboptimalno. Z optimizacijo usmerjanja API klicev bi lahko industrija prihranila do 70 % pri stroških inference. Kako?

AI Ecosystem Architecture with OpenRouter



Platforma **OpenRouter** deluje kot **univerzalni vmesnik (multiplexer)** in ponuja popoln nadzorni prerez trga, saj agregira podatke obeh vrst modelov.

Sistemska pravila: Zaprti proti Odprtim modelom

	Zaprti modeli	Odprti modeli
Dostop	 Proprietarni API, črna skrinjica.	 Hugging Face, javne uteži.
Infrastruktura	 Prvoosebna (1st-party) in ekskluzivni oblaki.	 Neodvisni ponudniki z optimizirano strojno opremo (3rd-party).
Cenovna dinamika	 Monopolne cene, visoke marže.	 Konkurenca znižuje ceno proti mejnemu strošku računske moči.

Sistemski paradoks (Analiza poletja 2025)

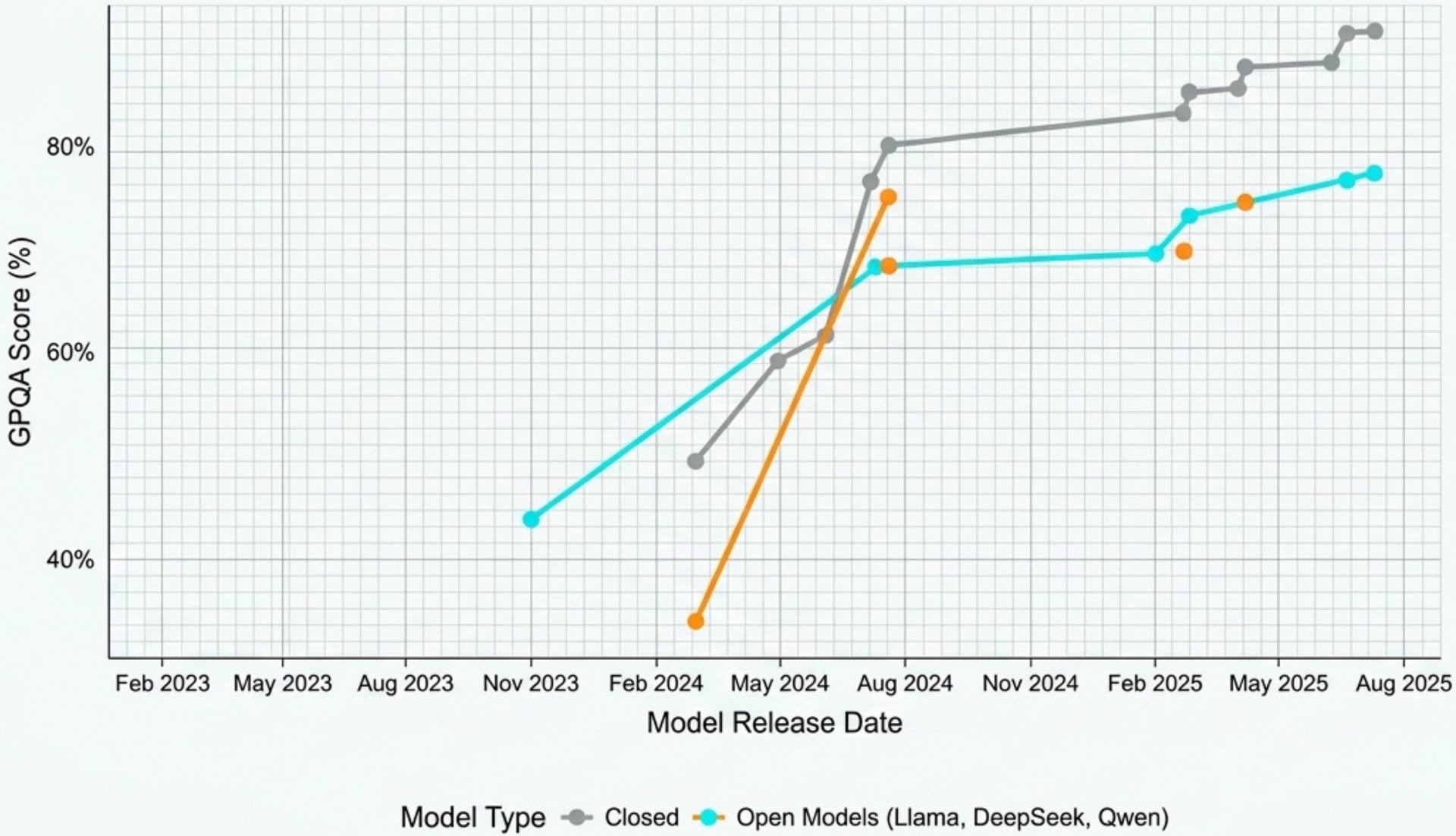
prihodkov si lastijo zaprti modeli.

zmogljivosti vodilnih modelov dosegajo odprti modeli.

nižja cena na uporabljen žeton pri odprtih modelih.

```
> POIZVEDBA: Zakaj inženirske ekipe plačujejo 6x več za samo 10 % izboljšanje zmogljivosti?
```

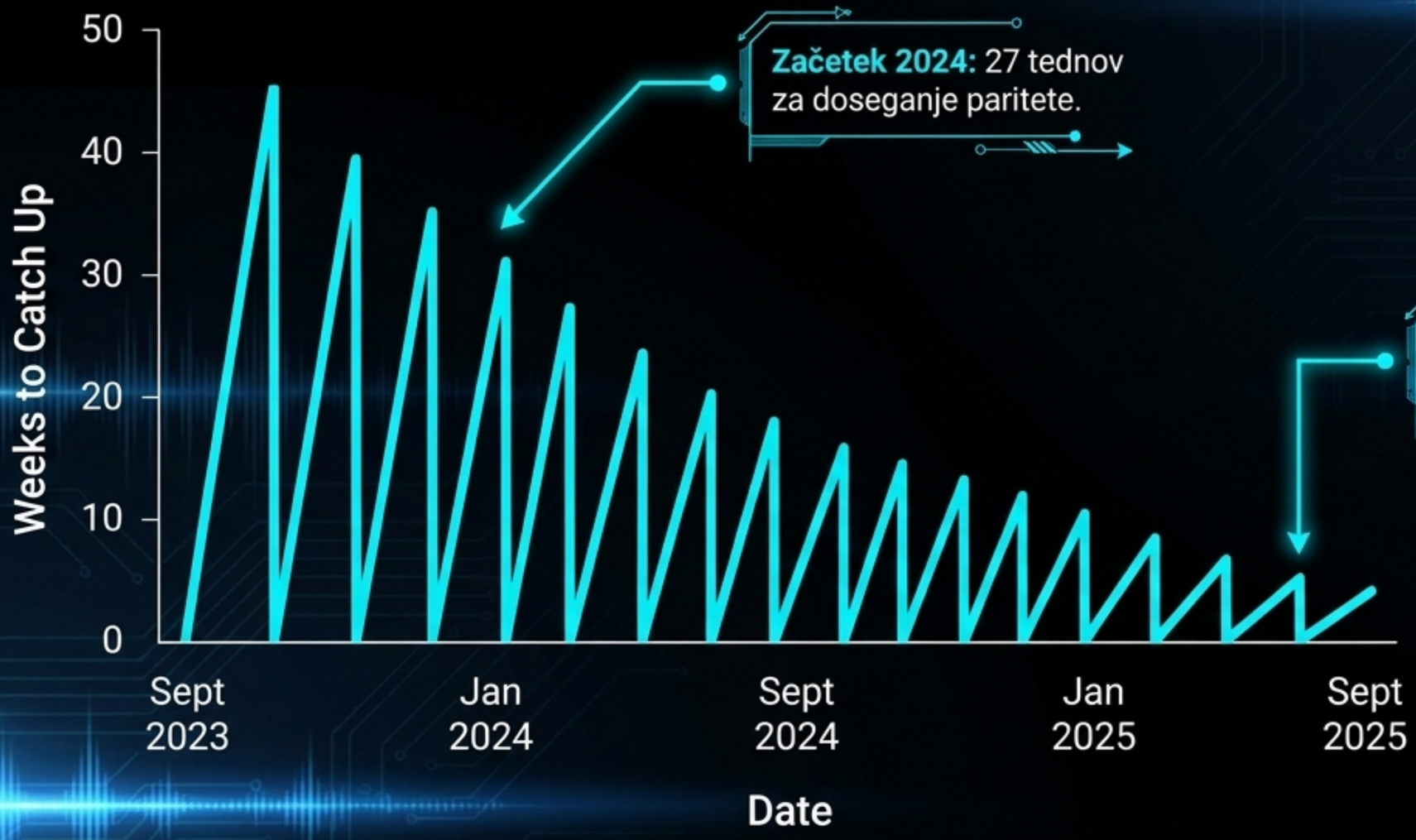
Iluzija zaščitnega jarka (The Moat Illusion)



Zaprte arhitekture ustvarjajo nove mejnike zmogljivosti, vendar odprti modeli (Llama, DeepSeek, Qwen) ta zaščitni jarek uničijo z izjemno hitrostjo.

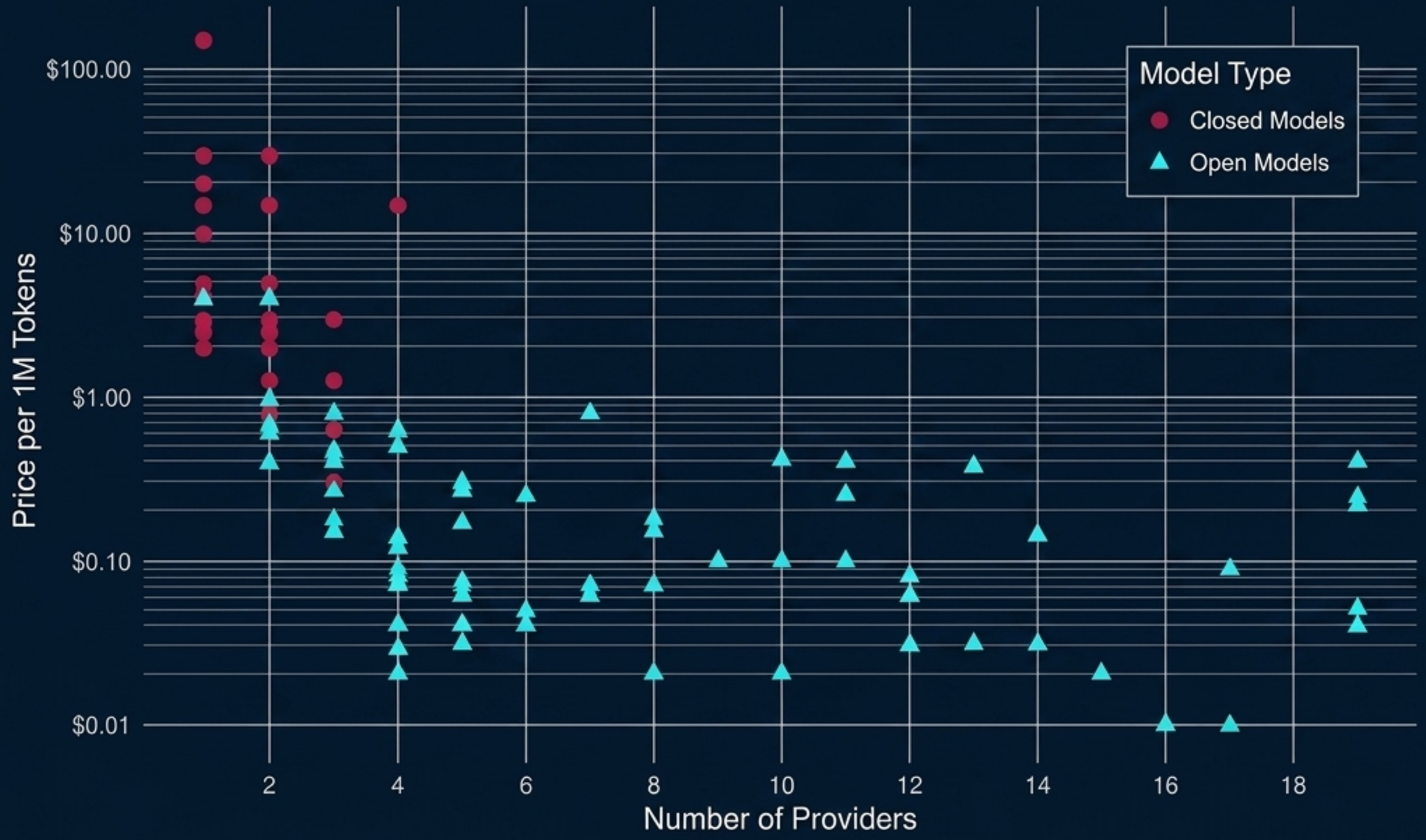
Tehnična prednost je le začasna.

Pospešena stopnja doseganja paritete



SKLEP: Hitrost dohitevanja raste. Možnost ohranjanja tržne prevlade zgolj na podlagi surove zmogljivosti modela se drastično zmanjšuje.

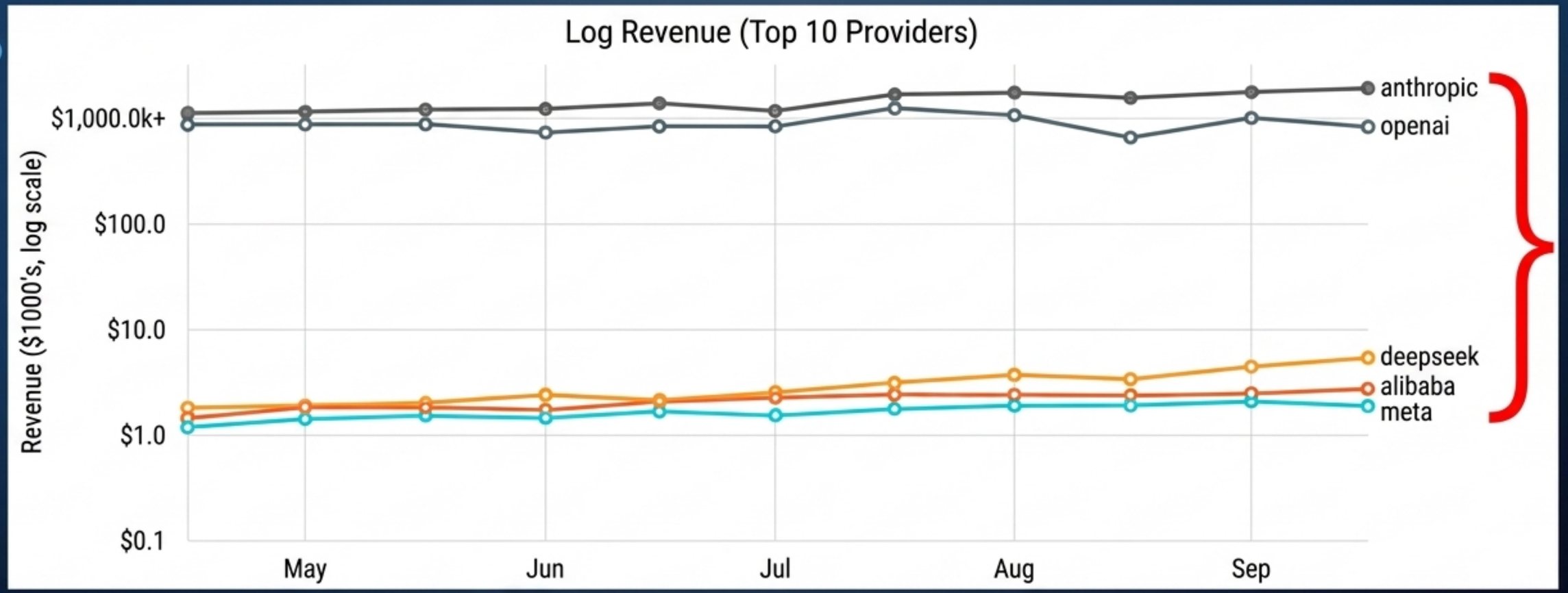
Fizika odprte konkurence



ZAPRTI MODELI:
Ohranjajo vertikalni monopol (1 ponudnik), kar omogoča visoke cene.

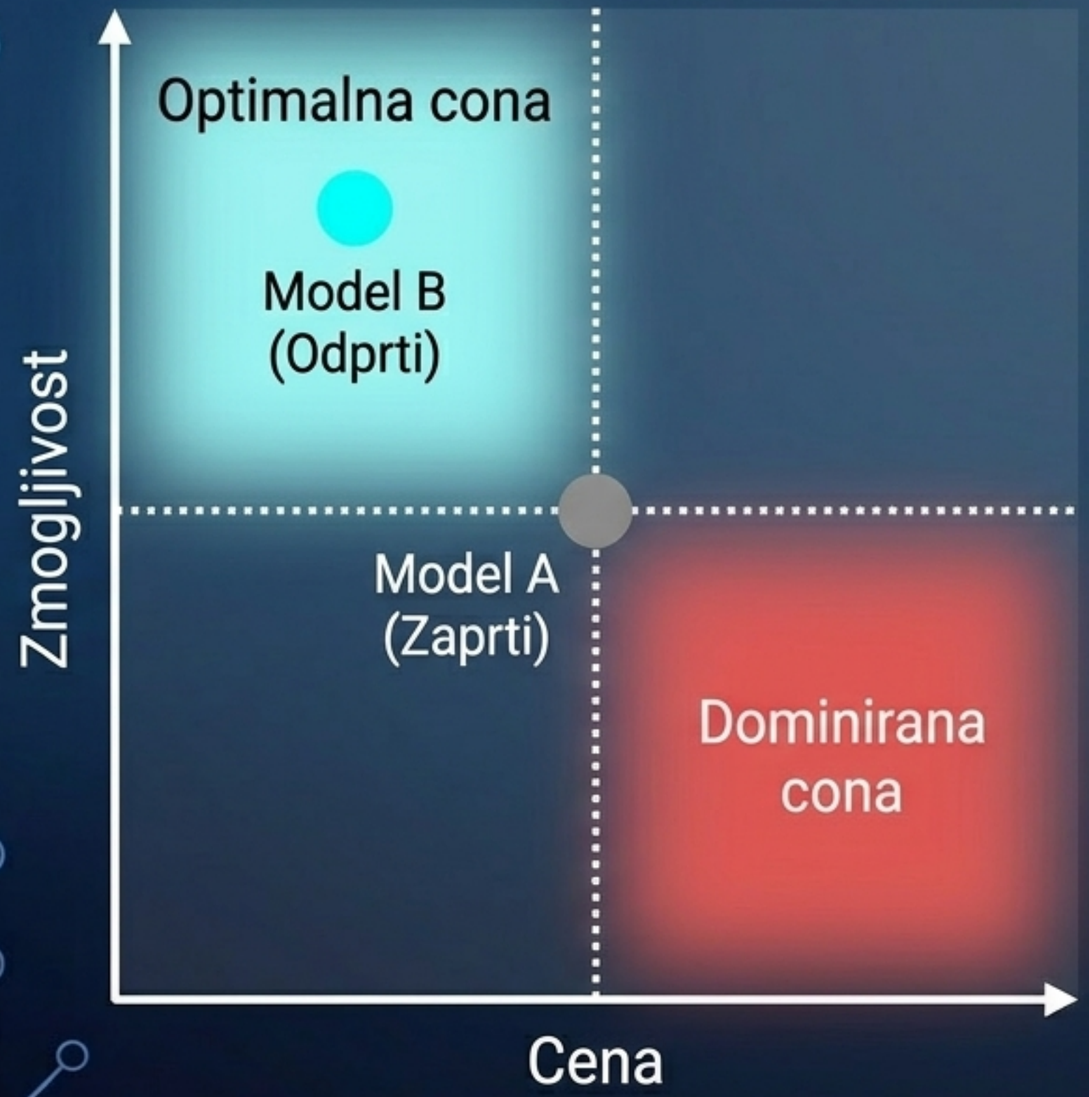
ODPRTI MODELI:
Več API ponudnikov (do 19) ustvarja strojno-optimizirano konkurenco, ki ceno znižuje proti dejanskemu strošku elektrike in silicija.

Diskonekcija pri uporabi (The Usage Disconnect)



ANOMALIJA: Čeprav matematika stroškov in zmogljivosti očitno favorizira odprte modele, inženirske ekipe v produkciji še vedno "trdo kodirajo" (hardcode) svoje aplikacije na drage, zaprte API-je.

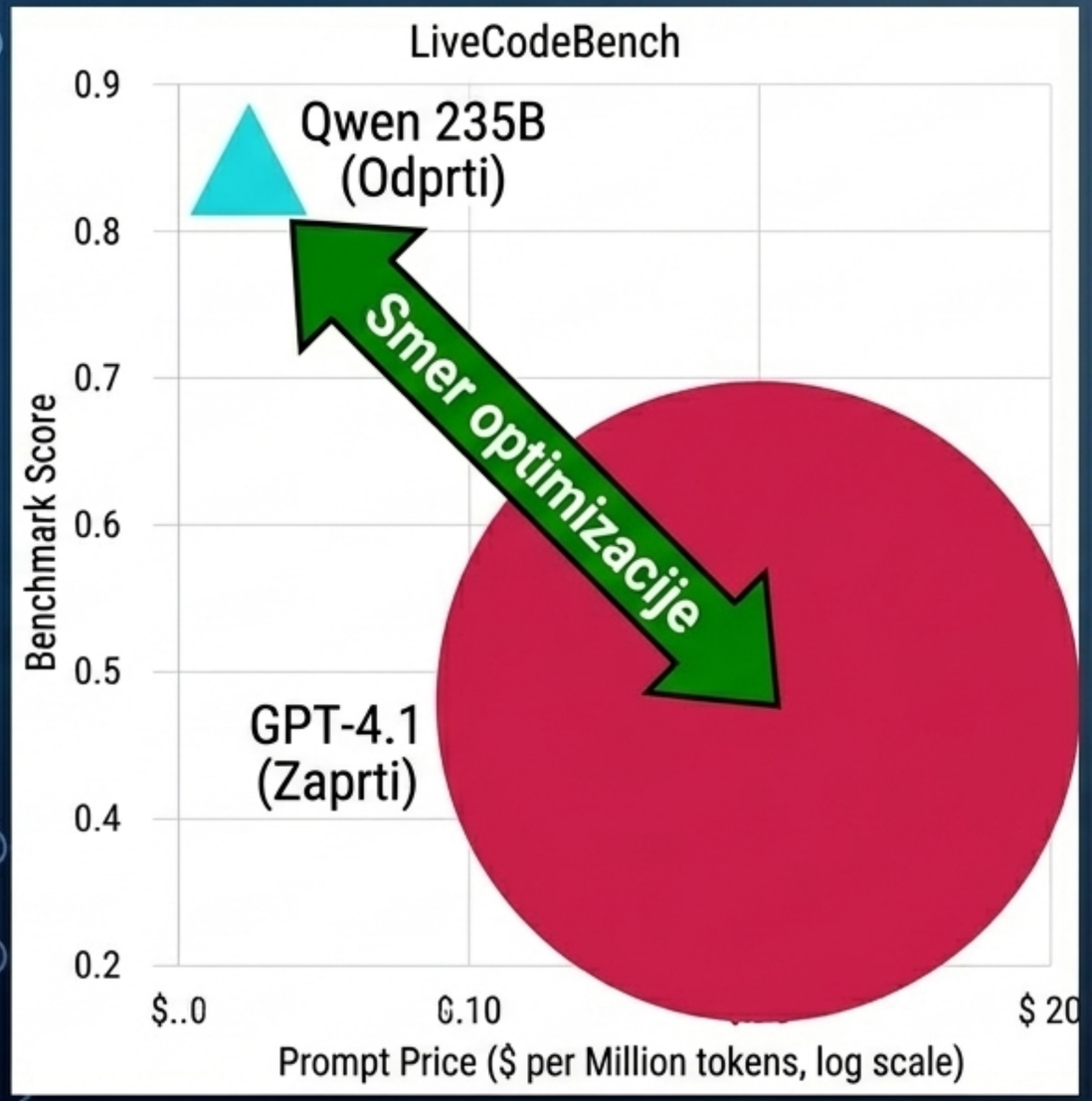
Definicija “Opazljive dominacije” (Pareto)




```
IF (Model_B.Zmogljivost >
    Model_A.Zmogljivost)
   AND
   (Model_B.Cena <
    Model_A.Cena)
THEN
   Model_A = DOMINIRAN
```

Če aplikacija uporablja model v rdeči coni namesto v zeleni, sistem porablja kapital za slabše rezultate.

Suboptimalni grozdi v produkciji



Model	Zmogljivost (GPQA)	Cena
GPT-4.1 (Zaprti)	0.67	$ 2.00
Qwen 235B (Odprti)	0.75	$ 0.18

ANOMALIJA: Velikost mehurčka (število porabljenih žetonov) pri modelu GPT-4.1 je kljub dominaciji Qwen 235B večstokrat večja.

Algoritem optimizacije (Simulacija)

```
# Analiza vseh API klicev poleti 2025

def optimize_routing(token_request):
    for model in available_models:

        # Preveri pogoj Pareto dominacije
        if (open_model.score > closed_model.score) and \
            (open_model.price < closed_model.price):

            # Izvedi optimalno preusmeritev
            route_to(open_model)
            calculate_savings()
```

Raziskovalci so ta algoritem counterfactual simulacije zagnali čez vse dejanske API klice uporabnikov platforme OpenRouter.

Tri strategije preusmerjanja

Vhodna poizvedba

1. Maksimizacija zmogljivosti

Izberi najboljši odprti model, ki je boljši od uporabljenega zaprtega (brez glede na dodatne prihranke).

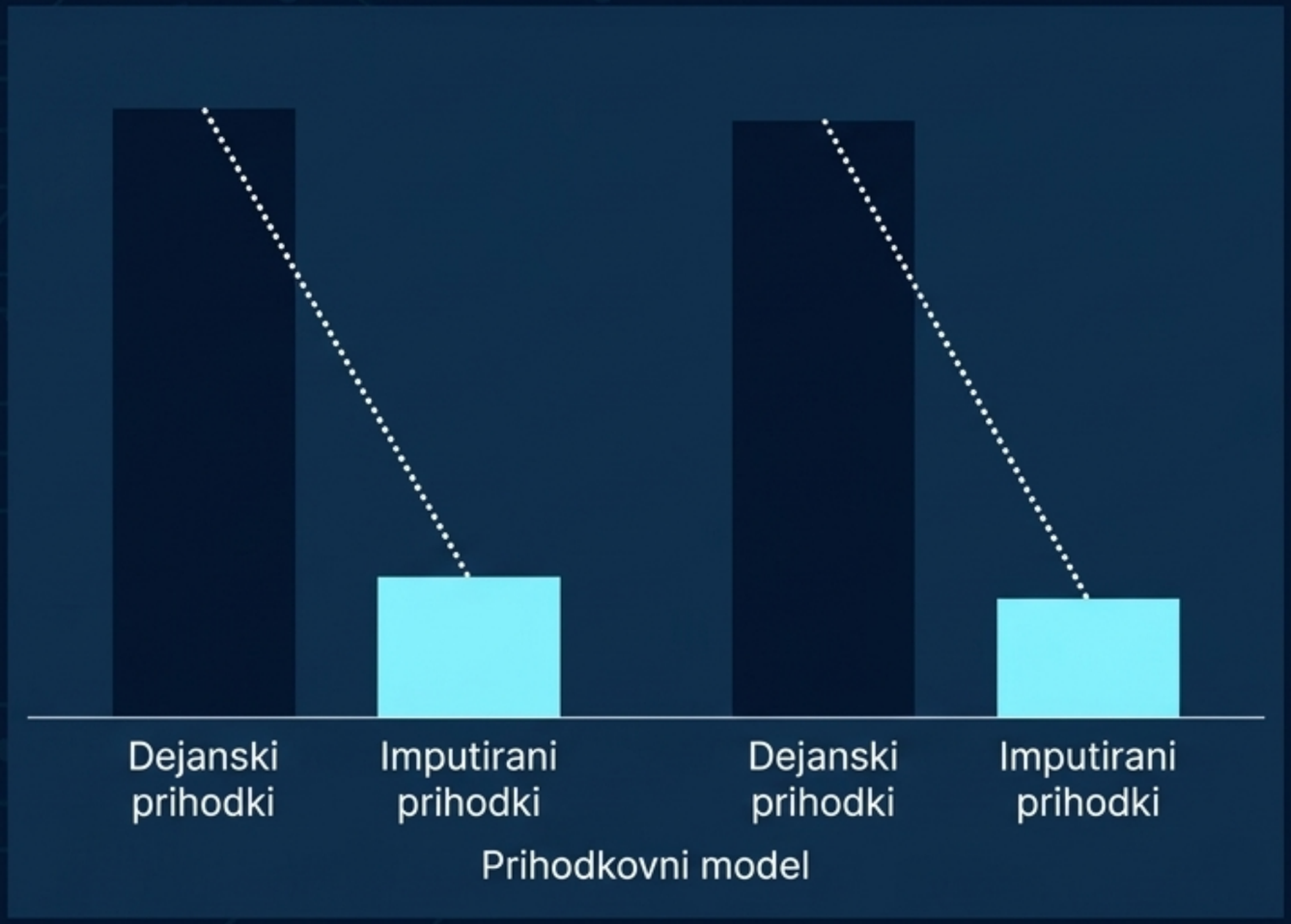
2. Minimizacija stroškov

Izberi najcenejši odprti model, ki še vedno prekaša trenutni zaprti model.

3. Najboljše razmerje

Izberi model, ki optimizira matriko cena-zmogljivost (Best Value).

Rezultati simulacije: Manjši stroški, boljši rezultati



~ 70.7 %

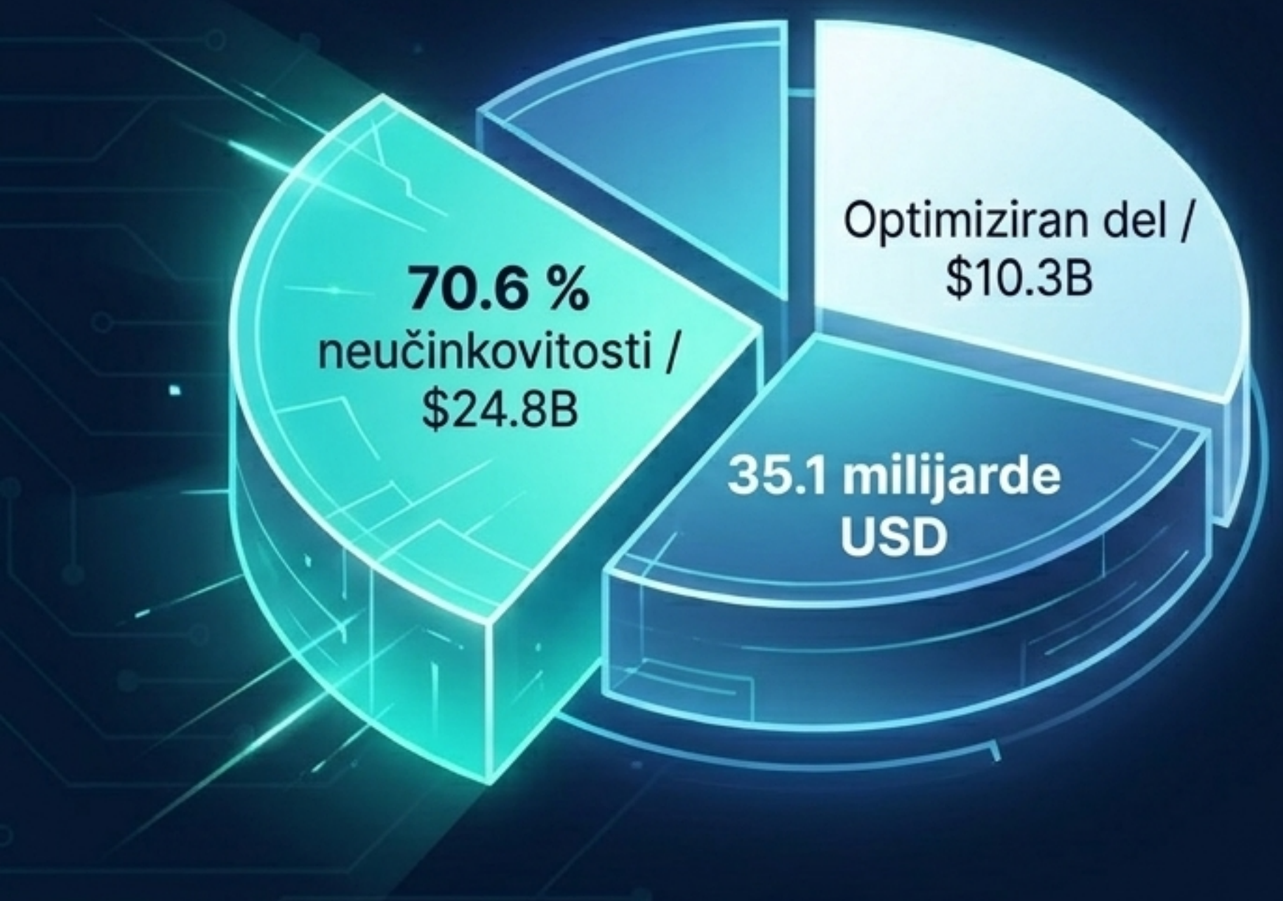
Padec povprečnih cen žetonov ob optimizaciji usmerjanja.

+ 14.3 %

Povišanje povprečne zmogljivosti na tehničnih testih.

Z uporabo preusmeritvenega algoritma bi uporabniki sočasno prihranili masivne količine denarja in izboljšali natančnost svojih aplikacij.

Ekstrapolacija makro neučinkovitosti



Skupni trg LLM inference (Ocena 2025)

35.1 milijarde USD

Povprečna stopnja podizkoriščenosti

70.6 %

IZRAČUN UHAJANJA KAPITALA:

24.8 milijard USD

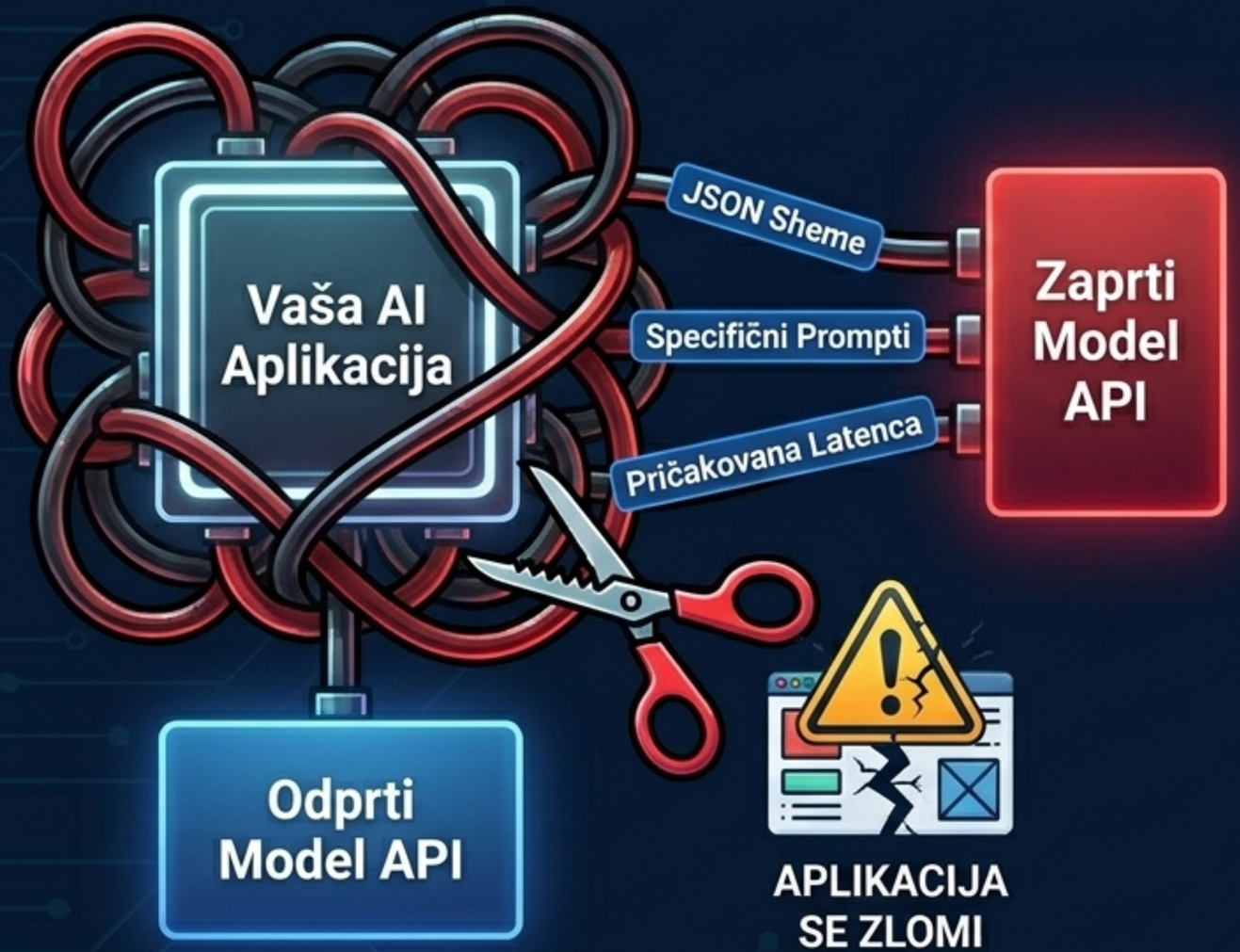
puščenih na mizi.

Industrija plačuje ‘davek na neoptimizacijo’ v višini skoraj 25 milijard dolarjev letno.

```
> SYSTEM.STATUS: Matematika je jasna.
> MARKET.STATUS: Trg se ne prilagaja.
>
> DIAGNOSTIKA: Iskanje neopaženih dejavnikov odločanja
(Unobserved Factors)...
> STATUS: Zagon razhroščevanja...
```

Zakaj inženirji ne zamenjajo API ključev,
ko so na voljo boljši in cenejši modeli?

Dejavnik 1: Arhitekturni ‘Lock-in’ in stroški zamenjave

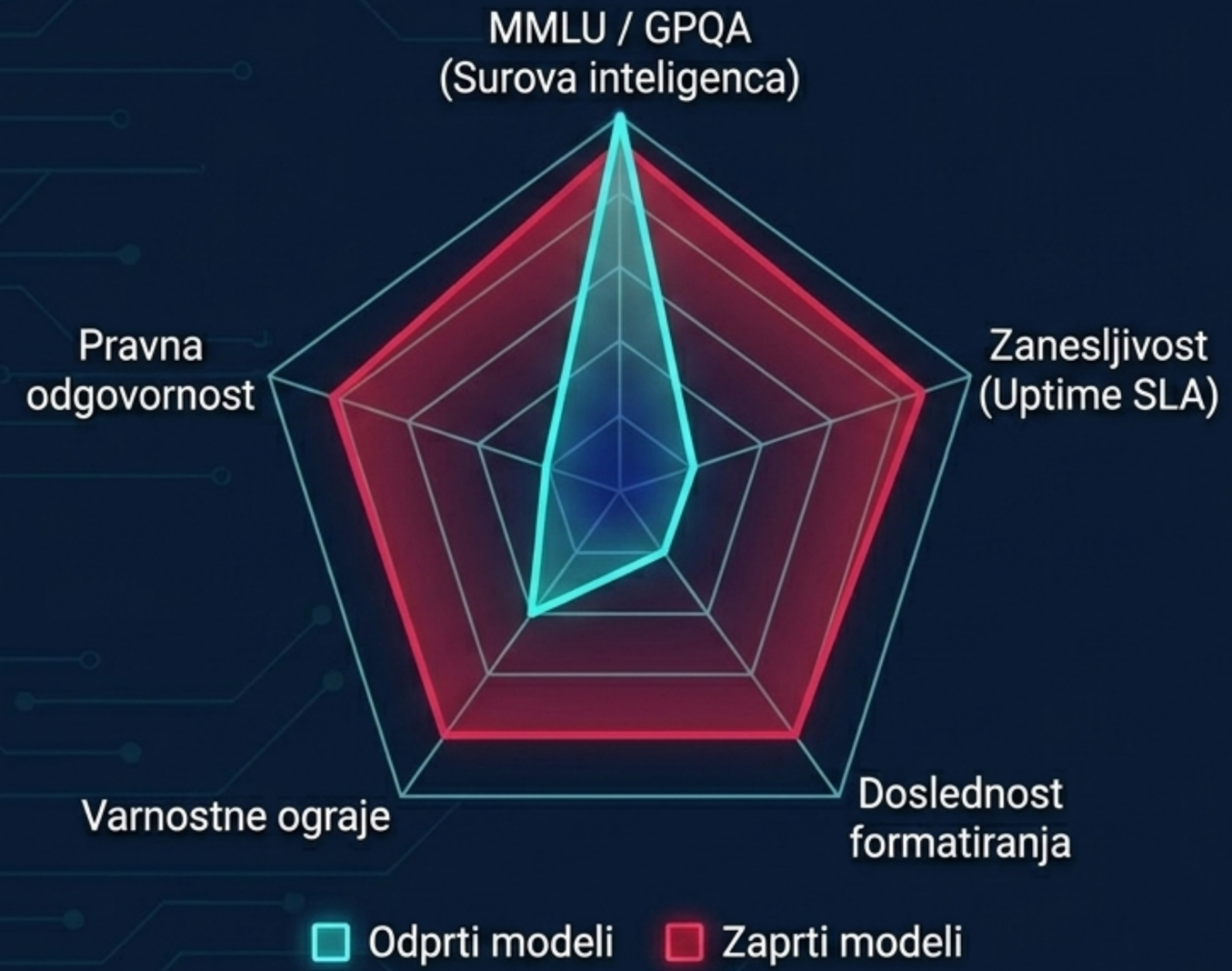


1. Idiosinkrazije promptov:
Aplikacije so ekstremno umerjene (“fine-tuned”) na specifično strukturo odgovorov določenega zaprtega modela.

2. JSON strukture:
Zamenjava modela pogosto zlomi kodo za razčlenjevanje (parsing) na nižjih nivojih aplikacije.

3. Latenca in časovni okviri:
Produkcijski sistemi so strogo umerjeni na specifične čase odgovorov prvega žetona (TTFT).

Dejavnik 2: Razkorak med testi in produkcijo



Standardizirani akademski testi (MMLU, GPQA, koda) merijo **zgolj surovo inteligenco**.

Kaj zahteva enterprise okolje:

- ☑ Zanesljivost in SLA dogovori (garancija delovanja)
- ☑ Doslednost strukturiranja izhodov v produkciji
- ☑ Varnostne ograje in zaščita podatkov
- ☑ Pravna in korporativna odgovornost ponudnika

Dejavnik 3: Institucionalna inercija

Nihče še ni bil odpuščen, ker je kupil OpenAI.

Psihološka varnost uveljavljenih blagovnih znamk ustvarja močno instituconalno zaščito.

Pogosto napačne predstave (strah, da odprti modeli pomenijo krajo podatkov) preprečujejo adopcijo.

Realizirana proti Nerealizirani vrednosti

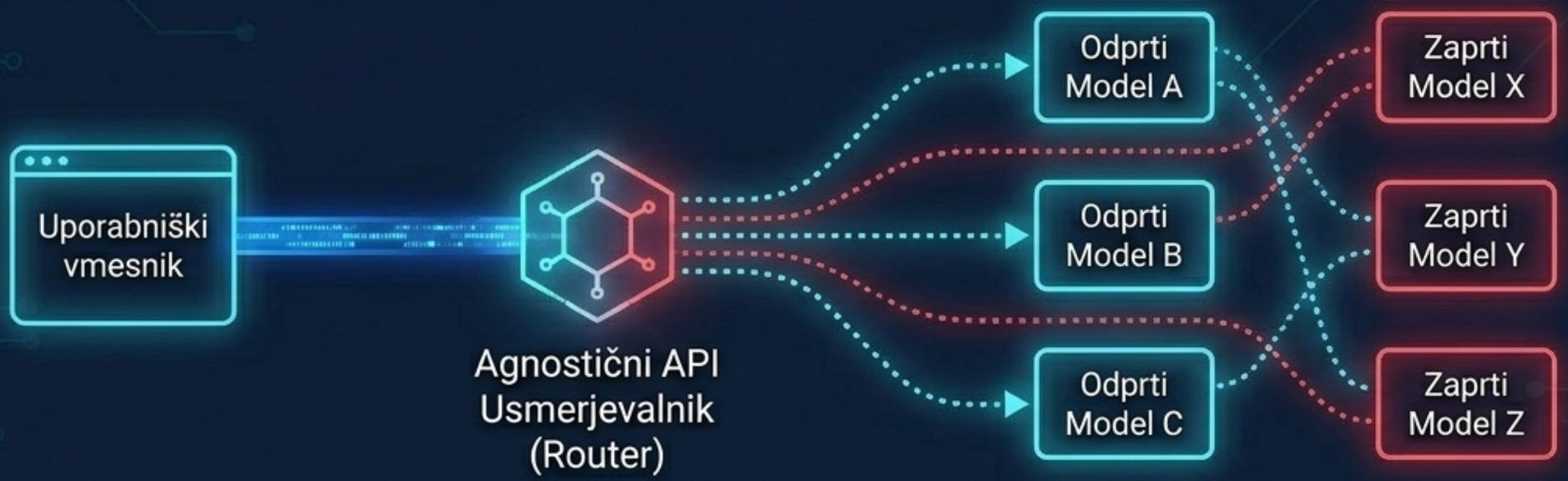
~1 milijarda USD

REALIZIRANO: Neposredni prihranki trenutnih uporabnikov odprtih modelov.

~25 milijard USD

NEREALIZIRANO (Latentno): Vrednost, ki jo je mogoče odkleniti zgolj s spremembo arhitekture usmerjanja.

Strateški zaključek za inženirje prihodnosti



1. Inferenca kot surovina (Commodity)

Ne obravnavajte ponudnika modela kot stalnega partnerja. Jezikovni modeli postajajo zamenljiva komponenta.

2. Arhitektura Multi-homing

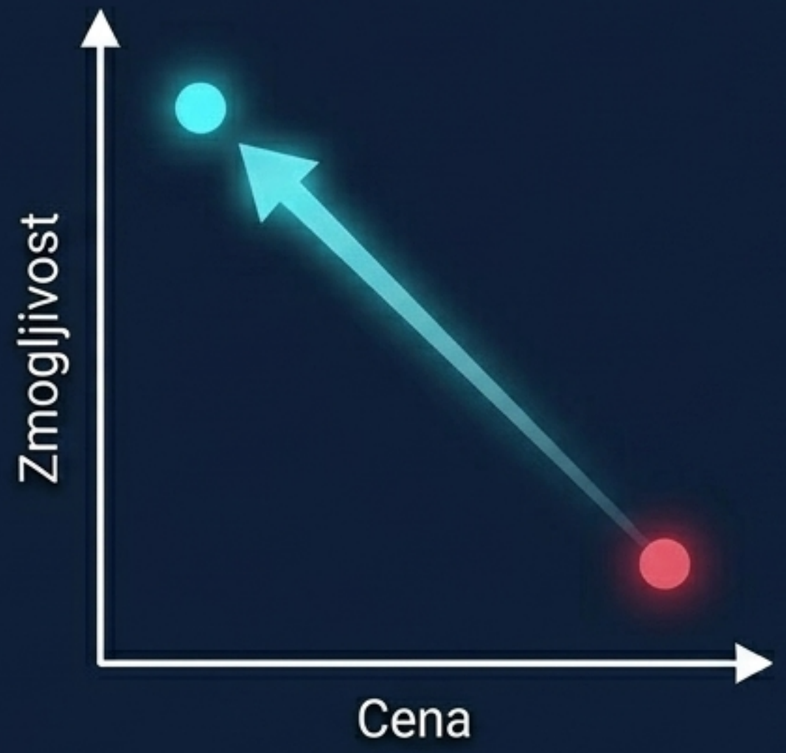
Aplikacije gradite preko abstraktnih usmerjevalnikov, ki omogočajo vročo zamenjavo API-jev brez spreminjanja izvorne kode.

3. Optimizacija na letenju

Vzpostavite lastno logiko ali uporabite platforme, ki usmerjajo klice na podlagi trenutne matrice cena/zmogljivost.

Končna sinteza: Prava vloga odprtih modelov

1. Hitri sledilci (Fast Followers): Tehnološki zaostanek se meri v tednih, ne več v letih.



2. Voditelji stroškovne učinkovitosti: Konkurenca na nivoju infrastrukture uničuje monopolne marže.

3. Zavarovanje trga (The Hedge): So ultimativni mehanizem proti oligopolnemu oblikovanju cen.

```
> Arhitektura umetne inteligence je problem optimizacije.
  Začnite jo tako tudi obravnavati.
```