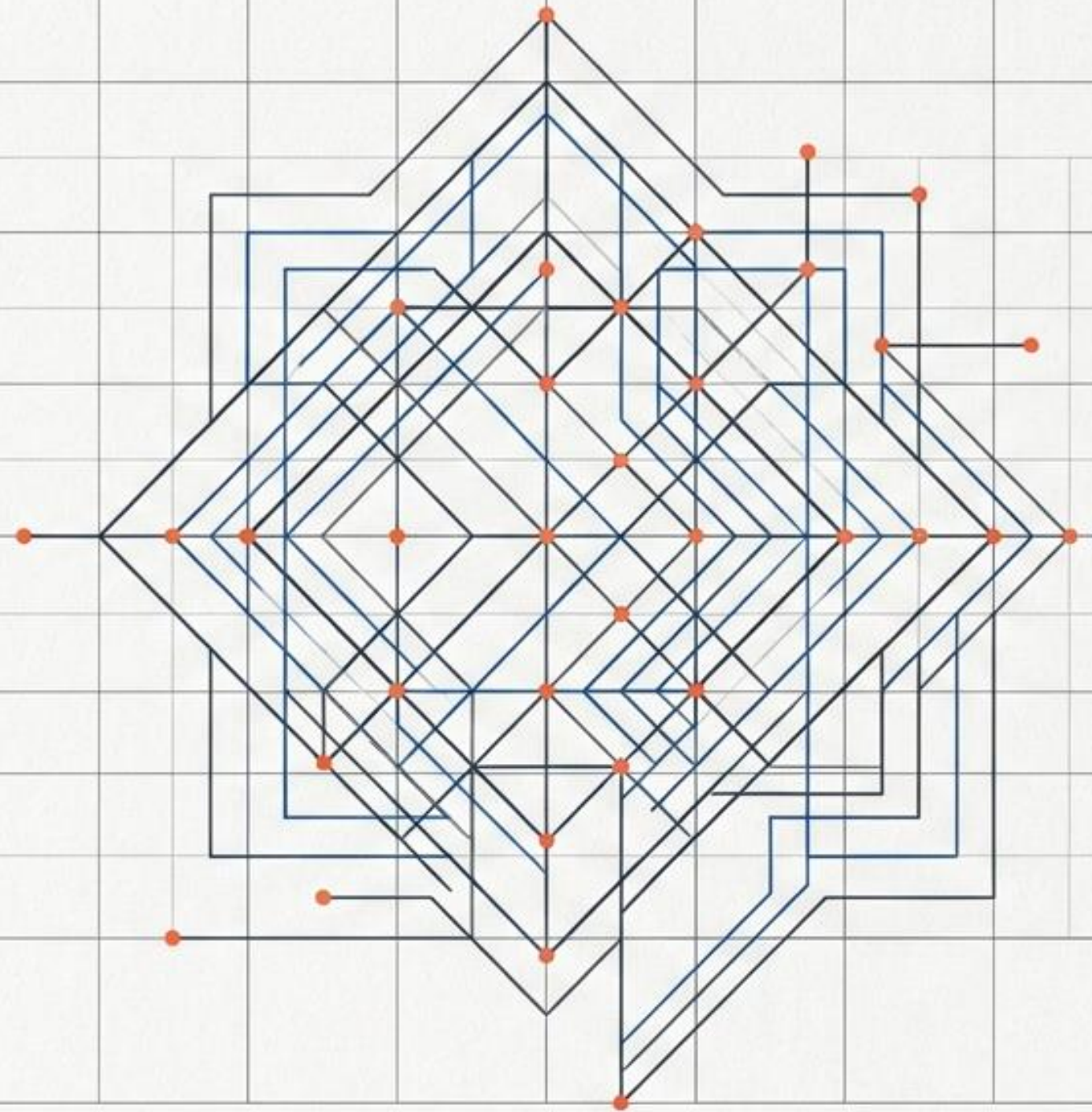


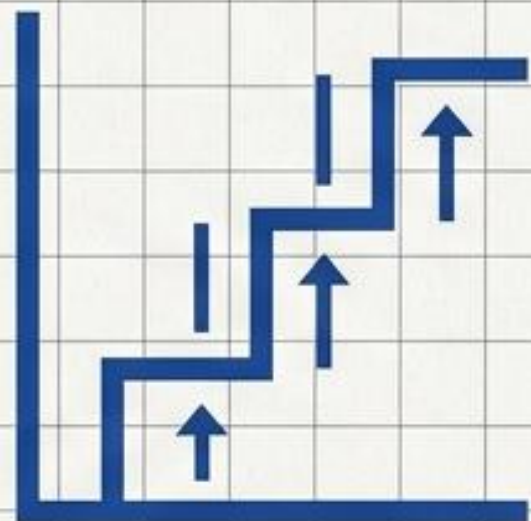
Meje umetne inteligence: Inženirski priročnik za izbiro modelov in API-jev

Zmogljivost, stroški in arhitekturna orodja
na podlagi telemetrije Artificial Analysis.

V4.3.2

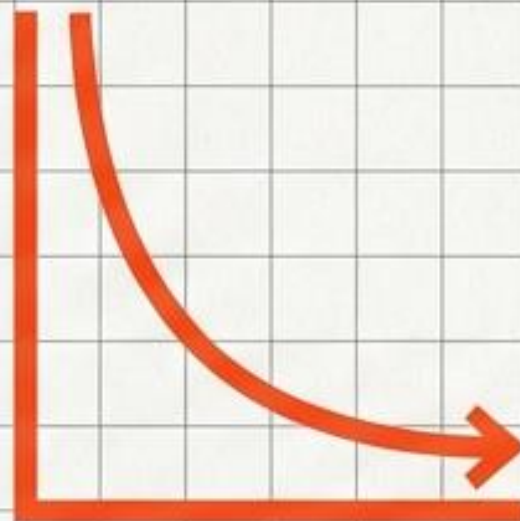


Stanje industrije: Ključni trendi



Plafon intelligence se viša

Modeli, kot sta Claude Opus 5.5 in GPT-6 Astra, premikajo meje zanesljivega sklepanja in avtonomnega reševanja inženirskih problemov.



Stroški eksponentno padajo

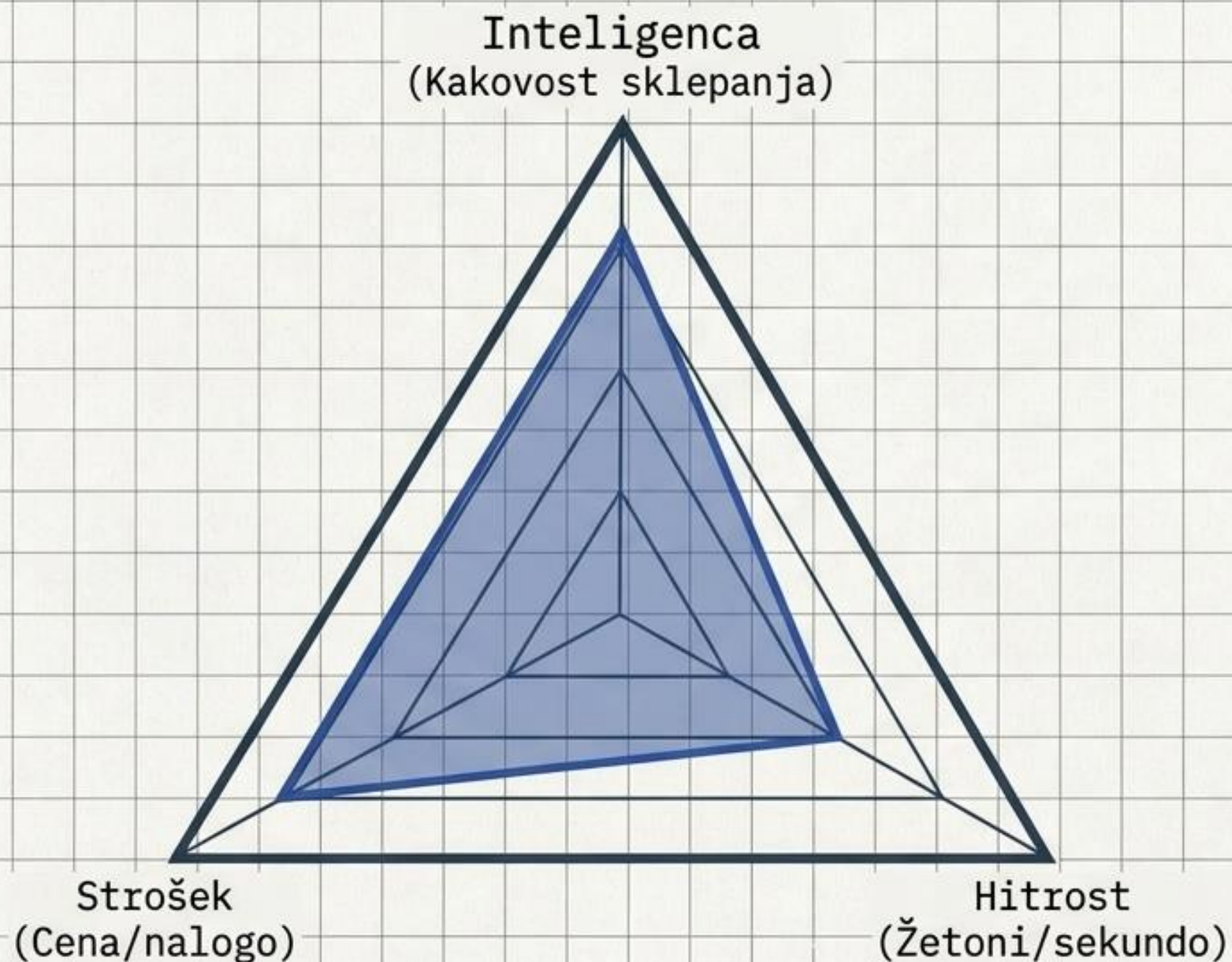
Visoko optimizirani arhitekturni modeli drastično znižujejo API stroške na opravilo in omogočajo ekonomsko upravičenost masovne obdelave.



Vzpon avtonomnih agentov

UI prehaja v integrirane agente (npr. Claude Code, Devin), ki samostojno izvajajo ukaze in rešujejo programske naloge v lokalnem okolju.

Inženirska trilema: Optimizacija kompromisov



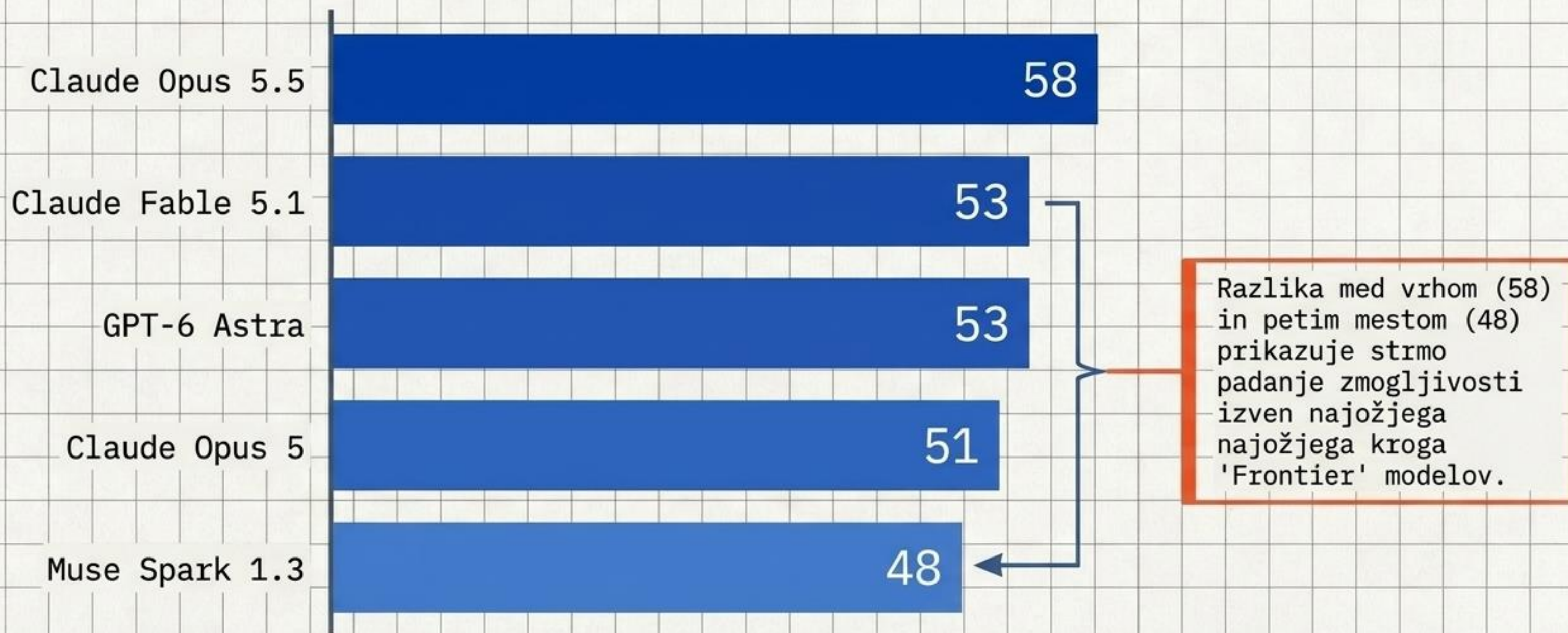
Univerzalni zmagovalec ne obstaja. Izbira API-ja je inženirski kompromis.

[Arhitektura: Real-time aplikacije]
-> Maksimizacija prepustnosti (Hitrost) ob sprejemljivi ceni.

[Arhitektura: Asinhroni Batch procesi]
-> Minimizacija stroškov za masovno obdelavo podatkov.

[Arhitektura: Kompleksni agentni sistemi]
-> Brezkompromisna zahteva po vrhunskem sklepanju (Inteligenca).

Indeks intelligence: Vrh absolutne zmogljivosti



Arhitekturna odločitev: Lastniški vs. Odprtokodni modeli

Oblak / API (Lastniški)

Najvišja absolutna inteligenca, zaprt sistem, plačilo po porabi.

Model	Score
Claude Opus 5.5	58
GPT-6 Astra	53

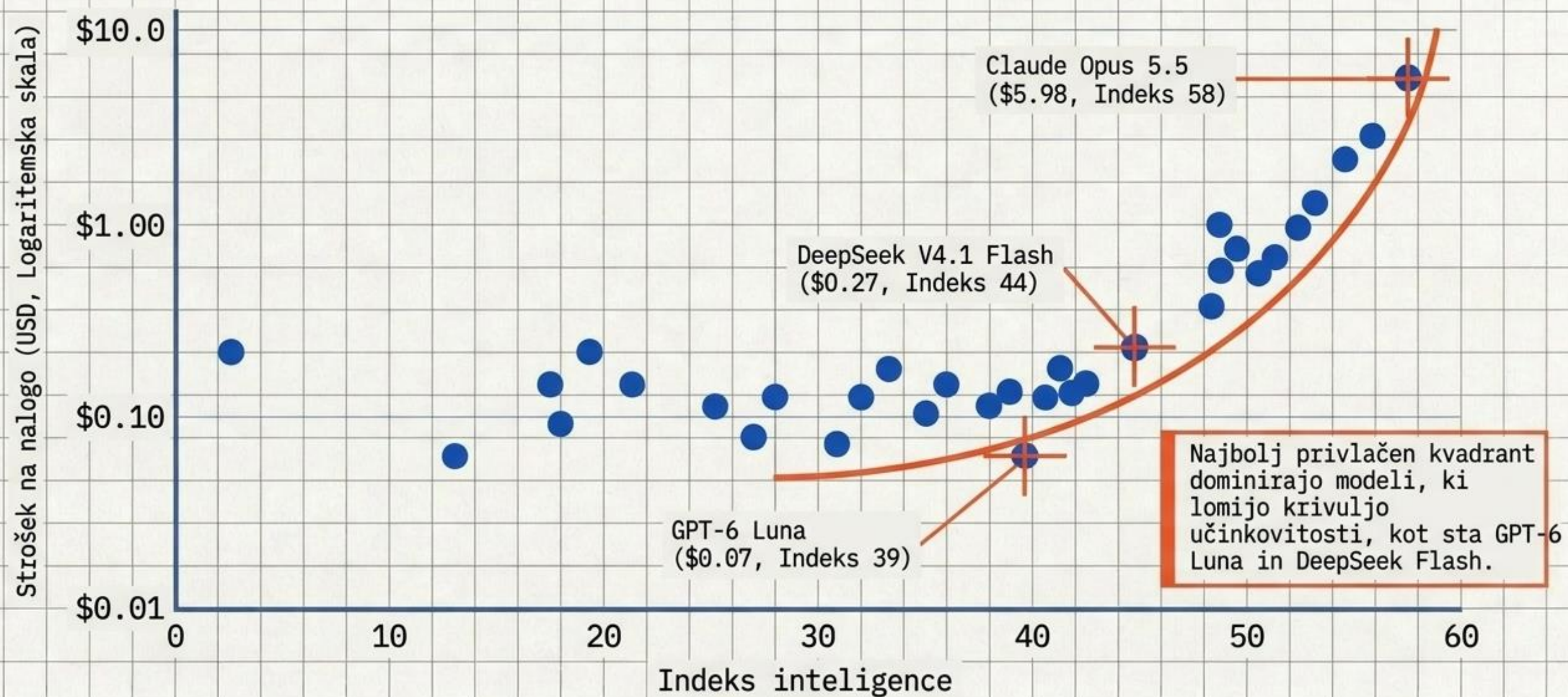
Lokalno / On-Premise (Odprtokodni)

Popoln nadzor nad podatki, možnost lokalnega izvajanja in prilagoditve (fine-tuning).

Model	Score
Qwen3.8 Max	47
DeepSeek V4.1 Flash	44
Nemotron 3 Ultra	25

Sinteza: Modeli kot sta Qwen in DeepSeek prinašajo izjemno zmogljivost v odprtokodni svet, kar omogoča gradnjo visoko zmogljivih in varnostno izoliranih internih sistemov.

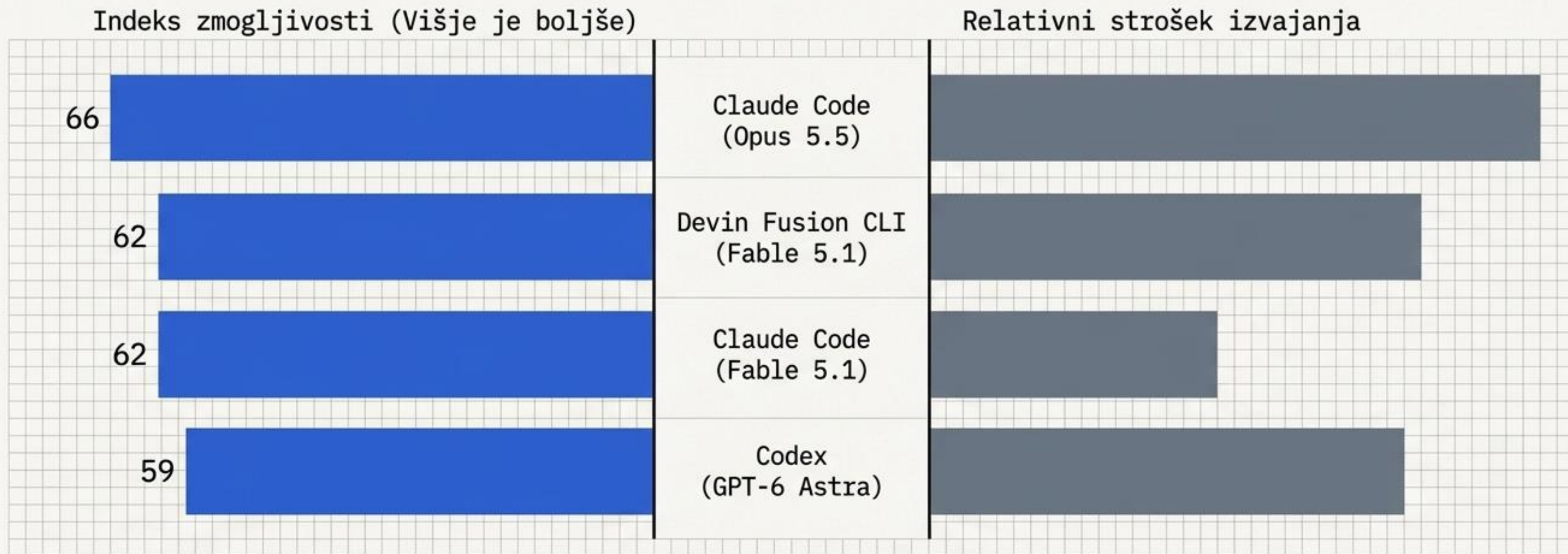
Optimizacija API stroškov in Pareto fronta



Prepustnost in zakasnitev za realnočasovne aplikacije



Agenti za kodiranje: Od LLM-ov do avtonomnih inženirjev



Konceptualni premik:

Agenti ne generirajo le kode; prevzamejo celoten življenjski cikel naloge v terminalu in IDE okolju, vključno z branjem repozitorija, zagonom testov in odpravljanjem napak.

Diagnostična matrika inženirskih zmogljivosti

	SciCode (Znanstveno programiranje)	Terminal-Bench 4.0 (Avtomatizacija)	CritPt (Fizikalno sklepanje)
Claude Opus 5.5	67%	60%	32%
Claude Fable 5.1	63%	52%	30%
GPT-6 Astra	56%	59%	32%
MiMo-V2.6-Pro	61%	35%	29%

Reševanje inženirskih problemov zahteva specializirano arhitekturo sklepanja. Zmagovalni modeli konsistentno izkazujejo superiornost pri nalogah, ki zahtevajo večstopenjsko logiko in interakcijo z zunanjimi orodji.

Sklepanje v dolgem kontekstu in nadzor halucinacij



Analiza dolgih
dokumentov
(AA-LCR v1.1)

- | | |
|-------------------|-------|
| 1. Kimi K3 | : 89% |
| 2. Step 5 Preview | : 88% |
| 3. MiMo-V2.6-Pro | : 86% |



Stopnja zanesljivosti
(Brez halucinacij)

- | | |
|--------------------|-------|
| 1. MiniMax-M3 | : 82% |
| 2. K2 Horizon 375B | : 74% |
| 3. GLM-5.3-Flash | : 72% |

Arhitekturni poudarek: Pri implementaciji RAG sistemov surova inteligenca ni dovolj. Zanesljivost sistema zahteva modele z visoko odpornostjo na halucinacije in natančnim priklicem iz ogromnih baz podatkov.

Večmodalnost: Vizualni modeli in razumevanje

Text-to-Image Leaderboard (Elo Rating)



GPT Image 2.5
Sunburst
1196 Elo



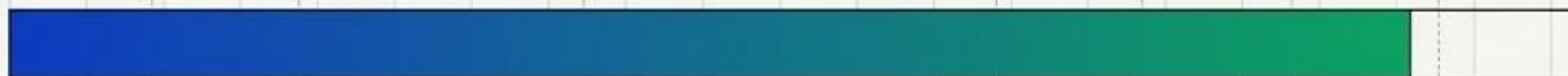
GPT Image 2.5
Flare
1190 Elo



Grok
Imagine 2.0
1154 Elo

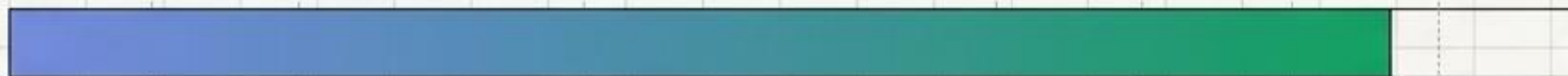
Vizualno sklepanje (MMMU-Pro)

Claude Opus 5.5



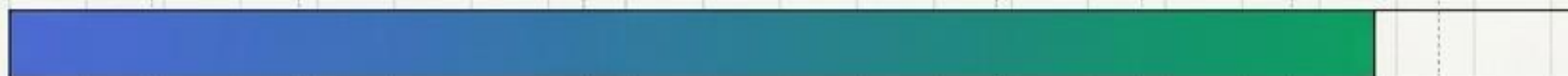
88%

GPT-6 Astra



87%

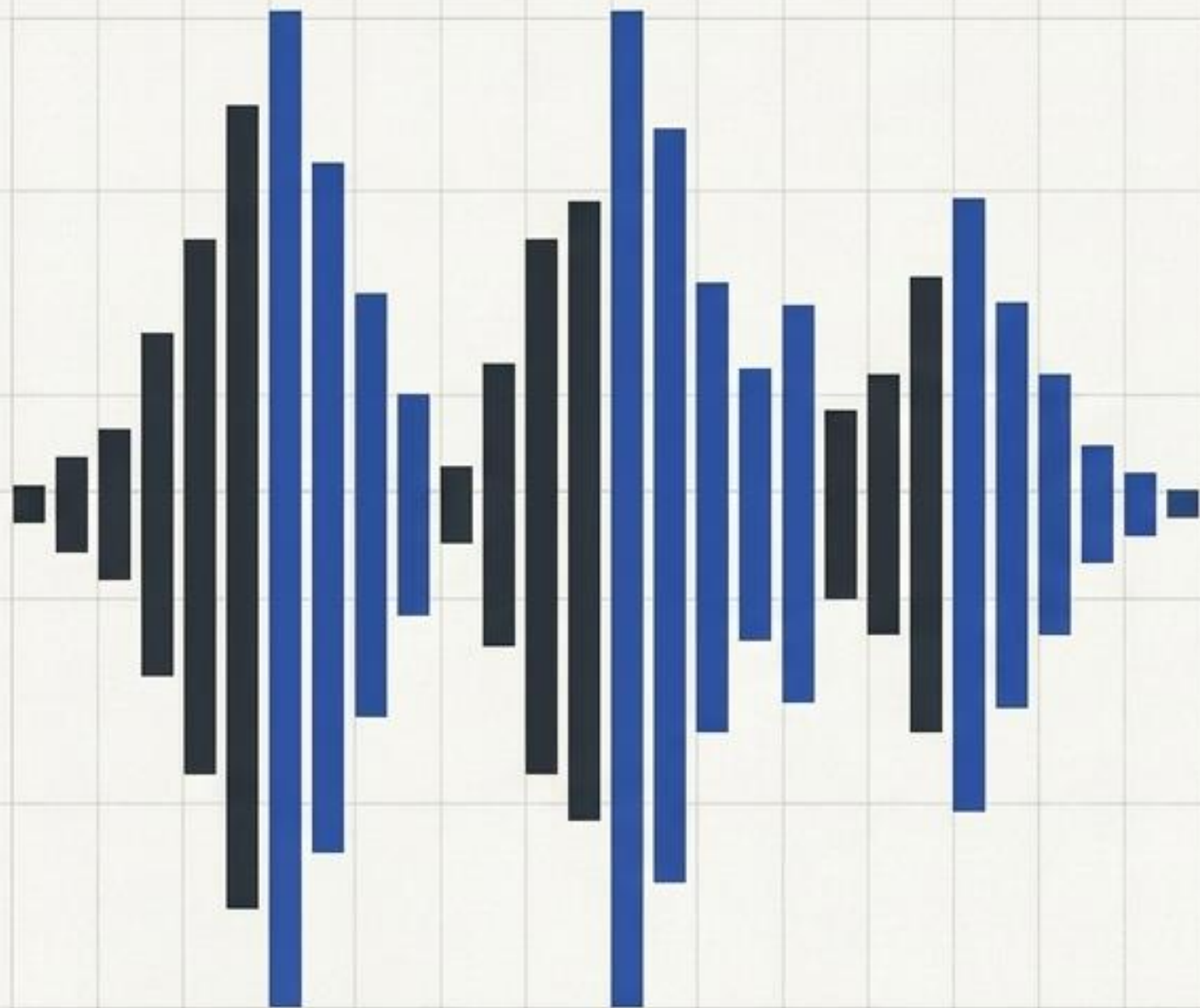
Gemini 3.8 Flash



86%

Generiranje in razumevanje slik prehaja k natančnemu razumevanju konteksta in tipografije, kar omogoča avtomatizacijo UI oblikovanja.

Glasovni vmesniki in interakcija človek-stroj (HCI)



Voice Arena Preference (Elo Rating)

1. Sonic 3.6	1273 Elo
2. Gemini 3.8 Flash TTS	1260 Elo
3. Qwen-Audio-3.0-TTS-Plus	1259 Elo
4. Realtime TTS-2	1245 Elo

Arhitekturni poudarek: Integracija Speech-to-Speech modelov z nizko zakasnitvijo omogoča gradnjo naravnih pogovornih vmesnikov v realnem času, ključno za robotiko in pametne asistente.

Odločitveno drevo za izbiro arhitekture in API-ja

Kateri model izbrati za tvoj projekt?

Zahtevaš vrhunsko
sklepanje ne glede na
ceno?

**Claude Opus 5.5 /
GPT-6 Astra**

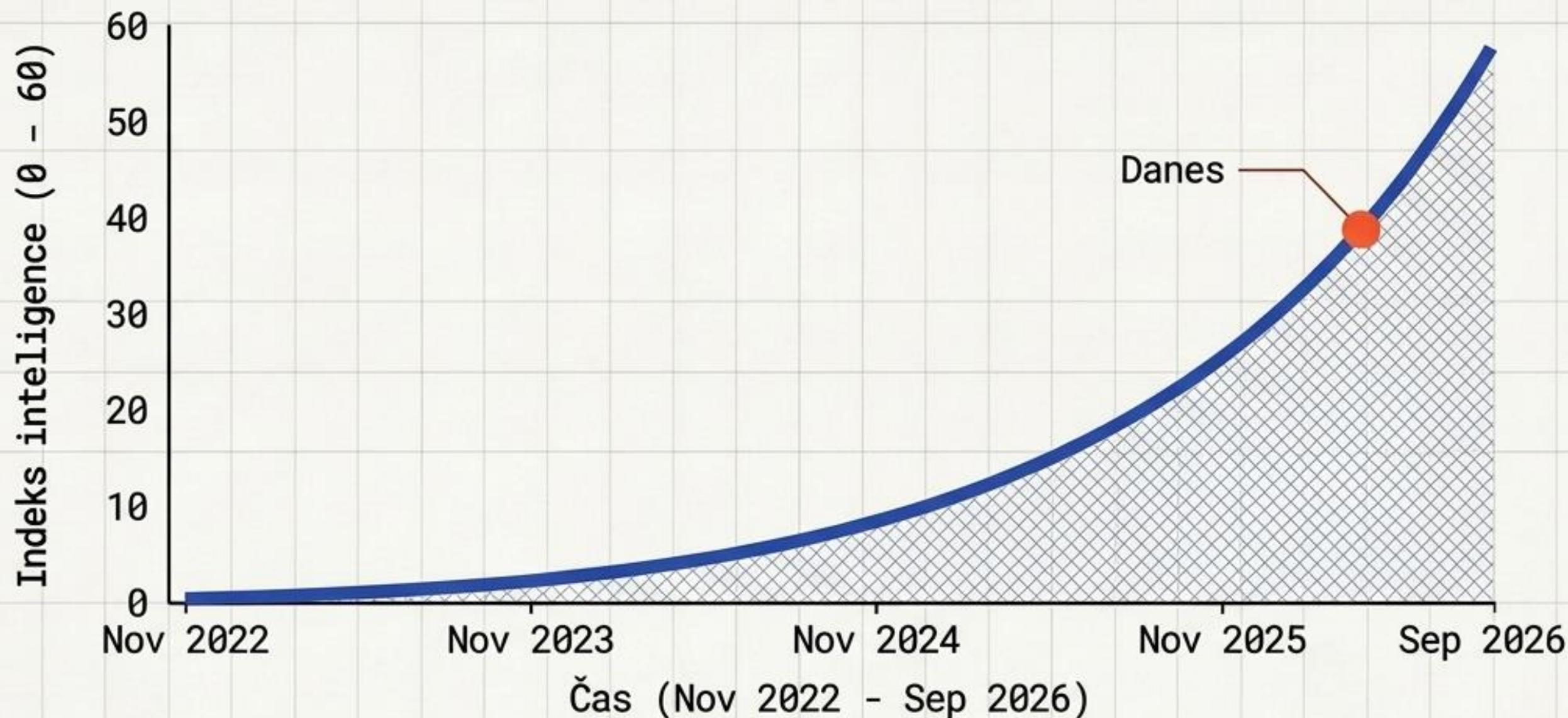
Iščeš optimalno razmerje
med hitrostjo, zmogljivo-
vostjo in nizko ceno?

**GPT-6 Luna /
Gemini 3.8 Flash**

Omejen proračun /
Ekstremna hitrost /
'Edge' naprave?

**DeepSeek V4.1 Flash
Qwen**

Evolucija intelligence in arhitekturna neodvisnost



Inženirsko pravilo: Tehnološka meja se nenehno premika. Sistemi morajo biti zgrajeni 'model-agnostic'. Vaša arhitektura mora omogočati hitro zamenjavo API vmesnikov, ko nova generacija modelov zlomi trenutno Pareto fronto.

Zaključek: The Architect's Checklist



Optimiziraj izbiro modela glede na specifično nalogo (Pareto fronta).



Uporabi odprtokodne modele za občutljive podatke in lokalno izvajanje.



Implementiraj avtonomne agente (Devin, Claude Code) za kompleksne operacije.



Nadzoruj stopnjo halucinacij pri RAG sistemih na velikih dokumentih.

Podatki: ArtificialAnalysis.ai