Deo 1: Ekosistem graničnih modela

Temelji nove inteligencije



Dobro došli u mašinsku sobu nove ekonomije. Ako je period od 2023. do 2025. godine bio obeležen „kambrijskom eksplozijom“ stotina različitih AI alata koji su često zbunjivali korisnike, 2026. godina donosi jasnu konsolidaciju i stratifikaciju. Više ne govorimo o apstraktnoj „veštačkoj inteligenciji“ kao monolitnom entitetu. Govorimo o **ekosistemu graničnih modela** (engl. *frontier models*), gde se nekoliko ključnih igrača izdvojilo ne samo po sirovoj snazi, već i po specifičnoj „filozofiji“ rešavanja problema.

U ovom delu knjige demistifikovaćemo tehnologiju koja pokreće ovaj talas. Nećemo se baviti matematikom iza neuronskih mreža, već funkcionalnom arhitekturom koja je relevantna za vas kao donosioca odluka. Razumećemo zašto jedan model košta pet dolara mesečno, a drugi pedeset, i zašto je onaj jeftiniji ponekad bolji izbor za vaš biznis. Istražićemo kako su se ovi modeli transformisali iz pasivnih generatora teksta u proaktivne agente sposobne da planiraju, vide, čuju i direktno upravljaju vašim računarom.

Za preduzeća na Balkanu, razumevanje ovog ekosistema je ključno za strategiju „preskoka“ (engl. *leapfrogging*). Nemamo resurse da pravimo sopstvene modele, ali imamo agilnost da biramo najbolje iz globalnog menija. U narednim poglavljima naučićete kako da taj meni čitate kao iskusni tehnološki somelijer – znajući tačno koji model upariti sa kojim poslovnim izazovom kako biste dobili maksimalnu vrednost uz minimalan trošak.

POGLAVLJE 1

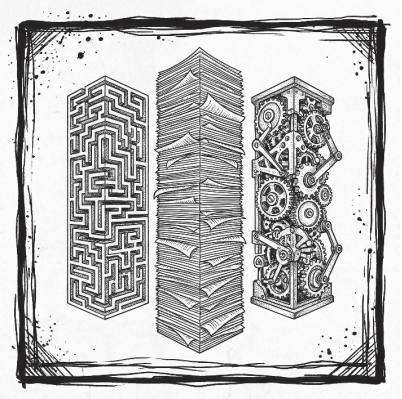
TRIOPOL 2026. GODINE

OpenAI, Google i Anthropic u post-GPT eri

KRAJ ERE „JEDAN ALAT ZA SVE“

Došli smo do trenutka zrelosti tržišta gde rečenica „koristim AI“ postaje jednako neprecizna kao rečenica „koristim softver“. U 2026. godini, pejzaž veštačke inteligencije kristalisao se u **triopol** – dominaciju tri tehnološka giganta koji definišu pravila igre, ali sa radikalno različitim pristupima: **OpenAI**, **Google** i **Anthropic**.

Za vas kao poslovnog korisnika, razumevanje razlika između njih nije stvar tehničke radoznalosti, već finansijske nužnosti i operativne efikasnosti. Svaki od ovih „velikih modela“ razvio je specifičnu „ličnost“ i set supermoći. Korišćenje pogrešnog modela za određeni zadatak je kao da pokušavate da zakucate ekser



šrafcigerom – moguće je, ali je sporo, frustrirajuće i rezultat je loš.

Ovo poglavlje je vaš taktički vodič za navigaciju kroz ovaj ekosistem. Naučićete kada da pozovete „stratega“ (OpenAI), kada „analitičara sa fotografskim pamćenjem“ (Google), a kada „pouzdanog izvršioca“ (Anthropic).

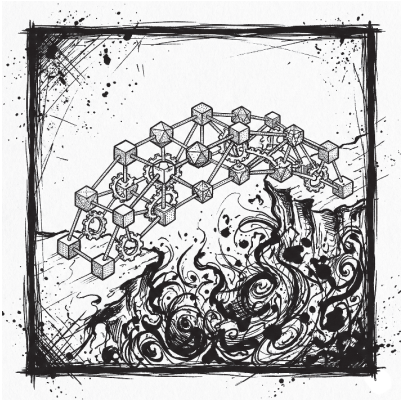
OPENAI: GOSPODAR REZONOVANJA I STRATEGIJE

Ako je GPT-4 bio trenutak kada je AI naučio da čita i piše na nivou prosečnog čoveka, serija modela **o1** i **o3** (koja je u 2026. postala industrijski standard) predstavlja trenutak kada je AI naučio da **razmišlja pre nego što progovori**. Ovo je fundamentalna promena u načinu na koji mašine obrađuju informacije, pomerajući se sa brze, intuitivne reakcije („Sistem 1“ razmišljanje) na sporu, logičku dedukciju („Sistem 2“).

Rezonovanje u vremenu izračunavanja

Za razliku od starijih modela koji su „izbacivali“ reči čim pritisnete Enter (poput automatskog dovršavanja rečenica na steroidima), modeli klase **o3** svesno prave pauzu. Oni koriste skriveni „lanac misli“ (engl. *chain-of-thought*) da bi interno debatovali,

simulirali različite pristupe, proverili sopstvene pretpostavke i ispravili logičke greške pre nego što vam isporuče konačan odgovor.[1] Ovo ih čini sporijim, ali drastično preciznijim u rešavanju problema koje niko ranije nije video.



Gde OpenAI dominira u 2026:

- **Složena strategija:** Planiranje ulaska na novo tržište gde postoji mnogo varijabli, nepoznanica i konkurentskih pritisaka.
- **STEM zadaci:** Napredna matematika, kodiranje kompleksnih sistema, naučna istraživanja i rešavanje problema koji zahtevaju višestepenu logiku.
- **Autonomno pregovaranje:** Situacije gde agent mora da predvidi poteze druge strane i prilagodi svoju taktiku u realnom vremenu.

Koristite modele serije „o“ (o1, o3) za zadatke gde je cena greške visoka, a brzina nije presudna. Nemojte ih koristiti za pisanje jednostavnih mejlova ili četovanje – to je kao da vozite Formulu 1 do prodavnice. Za te potrebe, OpenAI nudi model **o4-mini**, koji je optimizovan za brzinu i nisku cenu, a i dalje zadržava solidne sposobnosti rezonovanja.[2]



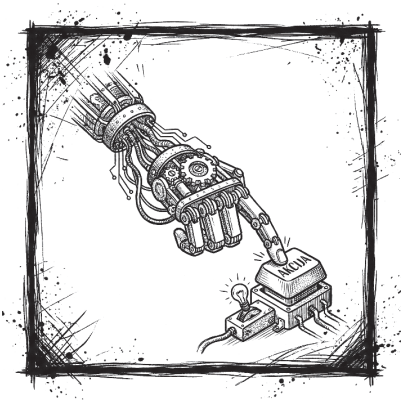
Optimizacija zaliha u srpskoj maloprodaji
Vodeći lanac sportske opreme u Srbiji koristio je standardni GPT-4 za predviđanje zaliha, ali rezultati su bili prosečni jer model nije mogao da poveže vremensku prognozu, lokalne praznike i istorijske podatke o prodaji u koherentnu logiku. Prelaskom na model **OpenAI o3**, implementirali su sistem koji „razmišlja“:

1. **Analizira:** „Najavljena je kiša za vikend u Beogradu, a u toku je školski raspust.“
2. **Rezonuje:** „To znači manju prodaju bicikala, ali veću prodaju kabanica i vodootporne obuće, posebno dečijih veličina.“
3. **Proverava:** „Da li imamo dovoljno zaliha u radnjama u centru i na Novom Beogradu?“
4. **Deluje:** Automatski generiše nalog za preraspodelu robe iz magacina u Krnješevcima ka ciljanim radnjama 48 sati pre promene vremena.

Rezultat: Smanjenje *out-of-stock* situacija za 35% u prvom kvartalu i povećanje margine profita na sezonskoj robi.[3]

ANTHROPIC (CLAUDE): EMPATIČNI INŽENJER I „COMPUTER USE“ REVOLUCIJA

Kompanija Anthropic, koju su osnovali bivši inženjeri OpenAI-a, zauzela je poziciju „etičkog i bezbednog“ lidera. Njihova serija modela **Claude** (posebno Opus 4.5 i Sonnet 4.5) u 2026. godini dominira u dva ključ-



na segmenta: **humanizovanoj komunikaciji** i direktnoj **kontroli računara**. Oni su shvatili da poslovni korisnici ne žele samo tekst, već žele da AI *uradi* posao u njihovim postojećim alatima.

Dok drugi modeli zahtevaju složene API integracije (povezivanje softvera kodom, što je skupo i sporo), Anthropic je pionir koncepta *Computer Use*. To znači da Claude može da „gleda" u ekran vašeg računara, pomera miša, klikće i kuca, baš kao što bi to radio čovek. Za srpska preduzeća koja često koriste zastarele softvere (knjigovodstveni programi iz 2010. godine koji nemaju API, državni portali), ovo je spasonosno rešenje.[4]



Automatizacija „pešačkog" posla u administraciji

Zamislite radnika u administraciji transportne firme koji svakog jutra mora da:

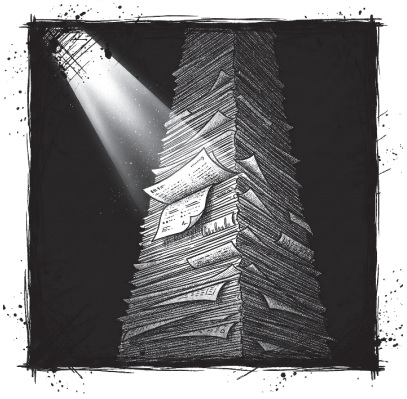
1. Otvori PDF fakturu koja je stigla mejlom od dobavljača goriva.
2. Kopira podatke (PIB, iznos, poziv na broj, litražu).
3. Otvori stari ERP softver (koji radi samo na Windows-u i nema cloud verziju).

4. Zalepi te podatke u odgovarajuća polja, izabere šifru troška i klikne „Knjiži“. **Rešenje:** Claude 4.5 agent može da preuzme ovaj posao *gledajući ekran* i izvršavajući klikove. Nije potrebno programiranje interfejsa, samo instrukcija na prirodnom jeziku: „Odradi knjiženje svih faktura iz foldera ‘Ulazni’ i javi mi ako je neki iznos sumnjivo visok“.

⚠ Iako je „Computer Use“ moćan, on je sporiji od API integracije i nosi rizik da agent „klikne pogrešno“ ako se prozor pomeri ili iskoči reklama. Uvek koristite ovaj pristup u kontrolisanom okruženju („sandbox“) pre nego što ga pustite na produkcioni sistem, i uvek zahtevajte ljudsku potvrdu za finansijske transakcije veće od određenog iznosa.

GOOGLE (GEMINI): MULTIMODALNI DIV SA BESKONAČNIM PAMĆENJEM

Google-ov adut u 2026. je **Gemini 3**, a njegova „ubitačna aplikacija“ (engl. *killer app*) je **kontekstualni prozor** (engl. *context window*). Gemini može da „drži u glavi“ **dva miliona tokena** informacija odjednom – to je ekvivalent stotinama knjiga, satima video zapisa ili celokupnoj bazi koda vaše firme.[5]



Ovo menja način na koji pristupamo podacima. Dok druge modele morate da hranite podacima na kašičicu (kroz kompleksne RAG sisteme), Gemini-ju možete da „bacite“ celokupnu arhivu poslovanja za prošlu godinu (PDF-ove, Eksel tabele, snimke sastanaka) i postavite kompleksno pitanje. Takođe, on je **nativno multimodalan**, što znači da nije treniran samo na tekstu, već razume video i zvuk kao svoj maternji jezik.



Da li vaša firma ima gomilu nesređenih podataka (stari ugovori, snimci Zoom sastanaka, tehnički crteži) iz kojih ne izvlačite vrednost jer niko nema vremena da ih pregleda? Ako je odgovor da, Google Gemini je verovatno najbolji alat za „rudarenje“ znanja iz te „mračne arhive“ (*dark data*).



Hibridni pristup

Najuspešnije kompanije u 2026. ne biraju pobednika, već kombinuju snage u takozvanim „složenim AI sistemima“ (*Compound AI Systems*).

- Koristite **Gemini** da pronađe relevantne informacije iz hiljada dokumenata (zbog velikog konteksta).
- Prosledite te probrane informacije **OpenAI o3** modelu da donese tešku stratešku odluku (zbog jačeg rezonovanja).
- Koristite **Claude** da napiše finalni, empatičan mejl klijentu (zbog boljeg tona i stila).

KOMPARATIVNA MATRICA ZA DONOŠENJE ODLUKA

Kako da znate koji alat da otvorite u ponedeljak ujutru? Koristite ovu matricu kao brz vodič.

Poslovni zadatak	Preporučeni model	Zašto?	Cena (relativna)
Pisanje osetljivog mejla / PR	Anthropic (Claude)	Najbolji osećaj za nijanse, ton i ljudsku psihologiju. Manje zvuči kao robot.	Srednja
Teška logika / Strategija / R&D	OpenAI (o1/o3)	Sposobnost „dubokog razmišljanja“ smanjuje halucinacije i greške u planiranju.	Visoka
Analiza velike arhive (Big Data)	Google (Gemini)	Ogroman prozor konteksta (2M+ tokena) omogućava analizu celih biblioteka odjednom.	Niska/ Srednja
Interakcija sa starim softverom	Anthropic (Computer Use)	Jedini model koji pouzdano koristi miša i tastaturu za navigaciju kroz GUI.	Visoka
Masovna automatizacija (jeftina)	GPT-4o mini / Gemini Flash	Izuzetno brzi i jeftini modeli za jednostavne zadatke (npr. ekstrakcija podataka iz računa).	Veoma niska

PROMPT ZA PRIMENU: STRATEŠKO PLANIRANJE SA „O3“

Evo kako da koristite najmoćniji model rezonovanja (OpenAI o3) da rešite realan poslovni problem. Primetite strukturu prompta: dajemo ulogu, kontekst, specifičan zadatak i ograničenja.

Prompt za primenu

Uloga: Ti si Chief Strategy Officer (CSO) za srednje preduzeće u industriji prerade hrane u Srbiji. Imaš iskustvo u kriznom menadžmentu i izvozu na EU tržište.

Kontekst: Naši troškovi sirovina (maline i šljive) su porasli za 15% zbog suše, a konkurencija iz EU (Poljska) spušta cene gotovih proizvoda. Imamo stabilan obrtni kapital, ali nam je oprema za pakovanje zastarela.

Zadatak: Koristeći svoje napredne sposobnosti rezonovanja, nemoj odmah dati rešenje. Prvo razmisli korak po korak (*think step-by-step*):

1. Analiziraj tri moguća scenarija za narednih 12 meseci (Pesimistički, Realni, Optimistički) uzimajući u obzir makroekonomske trendove.
2. Identifikuj skrivene rizike u strategiji „smanjenja cena“ da bismo pratili konkurenciju (*race to the bottom*).
3. Predloži radikalan pivot (zaokret) koji iskorišćava našu lokalnu poziciju i brend „Made in Serbia“, a koji konkurencija iz EU ne može lako da kopira.

Ograničenje: Rešenja moraju biti primenjiva u roku od 6 meseci i ne smeju zahtevati zaduživanje veće od 50.000 EUR.



ZAPAMTITE

- **Inteligencija je roba, rezonovanje je premija:** Cena generisanja običnog teksta pada ka nuli, ali cena kvalitetnog AI *razmišljanja* (planning/reasoning) ostaje visoka i vredna plaćanja.
- **Kontekst je kralj:** Koji god model da koristite, on je onoliko pametan koliko su kvalitetni podaci koje mu date. Google Gemini je beskoristan ako su vaši podaci zaključani u fioci ili na papiru.
- **Ne plašite se hibrida:** Vaša „AI radna snaga" treba da liči na pravi tim – imate stratega (OpenAI), istraživača (Google) i operativca (Anthropic). Nemojte forsirati jedan alat za sve.



VEŽBE

#1 REVIZIJA ALATA

Napravite listu od 3 procesa u vašoj firmi koji su „zapeli“ jer softver ne može da se poveže (npr. prebacivanje podataka iz Excela u e-banking ili državni portal). Proverite da li bi **Claude Computer Use** mogao to da reši.

#2 TEST REZONOVANJA

Uzmite jedan stari strateški plan ili budžet (iz prošle ili pretprošle godine). Ubacite ga u **OpenAI o1/ o3** (uz anonimizaciju podataka) i tražite da pronađe „logičke rupe“ koje ste tada propustili. Uporedite njegove uvide sa onim što se stvarno desilo na tržištu.

#3 „RAZGOVOR“ SA DOKUMENTOM

Uzmite dugačak, dosadan pravilnik (PDF) ili ugovor od 50 strana. Ubacite ga u Google Gemini ili ChatGPT. Postavite pitanje: „Koje su tri najvažnije obaveze za mene kao zaposlenog prema ovom dokumentu, i gde se kriju potencijalne kazne?“ Vidite da li je AI pronašao „sitna slova“ koja ste vi propustili.