KOFETKE.

50 / 50

VODNIKI O AI

Farma agentov z dejanskim proračunom delovanja

Kako čakalne vrste nalog, stroški modelov, nadzor in deleži neuspehov oblikujejo ekonomiko delovanja?



PRORAČUN DELOVANJA

Stroškovni model, ki temelji na sprejetem delu, ne na vpadljivem številu opravljenih korakov.

SLOVENSKA IZDAJA

september 2026
Praktična uporaba AI

Viri preverjeni 25. septembra 2026.

Dejavni agenti zlahka ustvarijo vtis, da ima sistem veliko zmogljivost za uporabno delo. Ena naloga lahko ustvari več klicev modela, več osnutkov in več sporočil ob pregledu, ne da bi nastalo eno samo sprejeto delo. Proračun delovanja potrebuje drugačen imenovalec: izhod, ki izpolni lastnikovo pravilo sprejema. Vključevati mora tudi stroške nalog, pri katerih tak izhod ne nastane.

Vodnika 18 in 19 sta obravnavala delitev dela in omejene urnike. Tu je tema ekonomika določene čakalne vrste: kaj vanjo vstopi, kaj porablja zmogljivost, kaj pregledovalec sprejme in zaradi česa lastnik delo ustavi. »Farma agentov« pomeni nadzorovano skupino izvajalnih procesov, ki prevzemajo naloge iz te vrste. Ne pomeni uvedenega podjetja ali neodvisnega vira prihodka.

Izberite enoto, ki jo je mogoče sprejeti

Predstavljajmo si izmišljeno servisno delavnico, ki iz poročil strank in odobrenega registra opreme pripravlja sprejemne pakete. Paket vsebuje pravilen identifikator opreme, kratek opis, povezan s predloženim poročilom, manjkajoče podatke in naslednje vprašanje za človeškega koordinatorja. Ne sme si izmisliti diagnoze, obljubiti datuma popravila ali poslati sporočila. Koordinator ga sprejme le, kadar so polja in sklici na vire uporabni za naslednji korak.

Delavnica v eno tedensko serijo sprejme 100 enoličnih sprejemnih nalog. Naloga ima eno identiteto, eno revizijo vhoda in en končni izid. Ponovni poskus modela je dodaten poskus pri

isti nalogi, ne nova zahteva stranke ali nov sprejeti izhod. Tudi urejene različice ostajajo znotraj izvirne naloge. Tako je imenovalec mogoče preveriti.

Podvojene dostave niso zgolj izmišljena težava čakalnih vrst. Amazon SQS dokumentira, da je kopijo sporočila iz standardne čakalne vrste mogoče prejeti znova, in priporoča idempotentne aplikacije.[3] To je lastnost navedene storitve, ne trditev, da vse čakalne vrste delujejo enako. V predlaganem sistemu delavnice identiteta naloge in revizija vhoda določata, ali sporočilo iz vrste nadaljuje obstoječe delo ali upravičeno začne novo. Štejte enolične sprejete naloge, hkrati pa ohranite strošek vsakega dejanskega poskusa.

Čakalna vrsta beleži tudi rok, pregledovalca, največje število poskusov in dovoljene vire podatkov. Nepopolni vhodi se vrnejo koordinatorju, še preden se začne delo modela. Izvajalni proces ne sme ugibati manjkajočih serijskih številk ali naloge narediti »dokončane« tako, da omili merila sprejema. Zavrnitev je viden izid, ne razlog, da nalogo izbrišemo iz stroškovnega poročila.

Pred izračunom cene uskladite serijo

Vse količine, cene in izidi v nadaljevanju so izmišljene predpostavke za načrtovanje, ne izmerjena uspešnost ali tarife ponudnikov. Serija ima 100 začetnih poskusov modela. Osemdeset jih pripravi tehnično popoln paket. Preostalih 20 dobi en dovoljen ponovni poskus; pri 12 ponovitvah nastane paket, osem pa jih še vedno ne uspe. Tako je poskusov modela 120 in paketov za presojo 92, različnih nalog pa še vedno le 100.

Koordinator na kratko pregleda vseh 100 končnih zapisov nalog, vključno z osmimi zapisi o neuspehu. Od 92 paketov jih 75 sprejme neposredno, 15 po dodatnem človeškem urejanju, dva pa zaradi vsebinskih pomanjkljivosti zavrne. Končni seštevek je 90 sprejetih in 10 nesprejetih nalog. Dodatno urejanje je človeško delo, ne dodatni klici modela.

Evidenca serije	Naloge ali poskusi	Kaj število pomeni
Začetne naloge in začetni poskusi	100	Vsaka sprejeta vhodna naloga se začne enkrat.
Dodatni poskusi modela	20	En ponovni poskus za vsako sprva neuspešno nalogo.
Tehnično popolni paketi	80 + 12 = 92	Paket obstaja; ni pa nujno že prestal pregleda.
Sprejeti paketi	75 + 15 = 90	Neposredni sprejem in sprejem po dodatnem urejanju.
Nesprejete naloge	8 + 2 = 10	Uporaben paket ni nastal ali pa je bil pri pregledu zavrnjen.

Ta razčlenitev je namenoma natančnejša od »90-odstotne uspešnosti«. Pojasni, kje prihaja do neuspehov in kaj porablja njihovo odpravljanje. Ponovni poskus lahko odpravi prehodno napako pri generiranju ali oblikovanju, ne more pa povrniti

podatka, ki ga v viru ni. Lastnik potrebuje oznake razlogov za 20 ponovnih poskusov in 10 končnih neuspehov, preden lahko presodi, ali bi pomagal drug model.

Vsak strošek štejte enkrat

Predpostavimo, da vsak poskus modela porabi 10.000 obračunanih vhodnih žetonov in 5.000 obračunanih izhodnih žetonov. Izmišljeni ceni API-ja sta 2 € na milijon vhodnih žetonov in 8 € na milijon izhodnih žetonov. Vhod vključuje navodila in predloženo gradivo; izhod zajema ves izhod modela, ki se obračuna po predpostavljeni tarifi, ne zgolj besedila, ki ga na koncu vidi koordinator. V tem načrtovalskem primeru se vsak poskus, tudi neuspešen, obračuna v celoti.

Strošek API-ja na poskus je **(10,000 ÷ 1,000,000 × €2) + (5,000 ÷ 1,000,000 × €8) = €0.06**. Strošek API-ja za serijo je **120 × €0.06 = €7.20**. Ne dodajte še ločenega »doplačila za ponovne poskuse« za 20 poskusov: ti so že vključeni v 120. V dejanskem preizkusu stalne količine žetonov zamenjajte z zapisi uporabe ter uskladite negotove ali pozno sporočene stroške, namesto da bi neuspešne klice šteli za brezplačne.

Prisotna je tudi lokalna predobdelava: 90 sekund aktivnega časa naprave na sprejeto vhodno nalogo oziroma 150 minut, kar ustreza 2,5 ure. Pri izmišljenem spremenljivem strošku 0,80 € na aktivno uro naprave to stane 2,00 €. Ta postavka pokriva le energijo med delovanjem in potrošni material. Razpoložljivost opreme je vključena v spodnjo razporeditev stalnih stroškov.

Predobdelava se v tem primeru ob ponovnih poskusih modela ne ponavlja; ponovna uporaba njenega shranjenega izhoda je del predpostavljenega delovnega toka.

Prevladuje človeški čas. Koordinator porabi tri minute za vsakega od 100 končnih zapisov, dodatnih šest minut za vsakega od 15 urejenih paketov in dodatne štiri minute za vsako od 10 nesprejetih nalog, da neuspeh zabeleži in preda v nadaljnjo obravnavo. Pri predpostavljeni vrednosti dela 30 € na uro te dejavnosti stanejo 150 €, 45 € in 20 €. Predaja neuspešne naloge ne pomeni, da je bil servisni sprejem ročno dokončan; morebitno poznejše dokončanje bi bilo ločeno delo z ločenim stroškom.

Strošek, vključen v to serijo	Izračun	Strošek
API, vključno s ponovnimi poskusi in neuspehi	120 × 0,06 €	7,20 €
Aktivna lokalna predobdelava	2,5 ure × 0,80 €	2,00 €
Začetni pregled končnih zapisov	300 minut ÷ 60 × 30 €	150,00 €
Dodatno človeško urejanje	90 minut ÷ 60 × 30 €	45,00 €
Dodatna predaja neuspešnih nalog	40 minut ÷ 60 × 30 €	20,00 €

Strošek, vključen v to serijo	Izračun	Strošek
Razporejeni ponavljajoči se stalni stroški	60 € na štiritedensko obdobje ÷ 4 serije	15,00 €
Razporejeni strošek začetne vzpostavitve	240 € ÷ 12 načrtovanih serij	20,00 €
Skupaj delovanje in delež vzpostavitve	Vsota teh postavk	259,20 €

Predpostavljeni stalni strošek 60 € vsebuje 40 € za programsko opremo, hrambo in razpoložljivost opreme ter 20 € za redno administracijo. Pregled servisnih nalog in predaja neuspehov v ta zneska nista vključena, ker že imata lastni postavki. To administracijo opravlja lastnik zunaj koordinatorjevega časa, namenjenega pregledom. Znesek 240 € za vzpostavitev pokriva začetno izvedbo in uvajanje zaposlenih. Razporeditev na 12 serij je notranji dogovor za vrednotenje, ne pravilo davčne amortizacije ali dokaz, da bo izvedenih vseh 12 serij.

To je ekonomski stroškovni model, ne zgolj napoved bančnega izpiska. Delo in lastna oprema lahko porabljata vire, ne da bi vsak teden nastal nov račun. Ponazoritev izključuje davke in financiranje ter predpostavlja, da ni drugih plačljivih orodij ali podatkovnih licenc. Dejanski lastnik mora dodati stroške, ki zanj veljajo. Obdobje stalnih stroškov je izrecno dolgo štiri tedne; to ni trditev, da ima vsak koledarski mesec štiri delovne tedne.

Delite s sprejetim delom

Rezultat je **€259.20 ÷ 90 = €2.88** na sprejeti paket. Če bi delili s 120 poskusi modela, bi dobili približno 2,16 € na poskus, kar odgovarja na drugo vprašanje. Cenejši poskus ni nujno cenejši sprejeti paket. Lahko povzroči več zavrnitev ali zahteva več urejanja.

Ohranite tako pregled serije kot pregled posamezne naloge. Pregled serije vključuje vse zavrnjeno delo in razporejene stroške. Pregled naloge pokaže, pri katerih vhodih so nastali ponovni poskusi in dolgi pregledi. Stroškov neuspešnih nalog ne prestavljajte v nevidno kategorijo »poskus«, medtem ko sprejete naloge predstavljate kot celotno storitev. Prav tako istega deleža stroška vzpostavitve ne prištevajte vsakemu posameznemu poskusu, potem ko ste ga že enkrat vključili v skupni strošek serije.

Za začasno primerjavo z ročnim delom predpostavimo, da ročna obdelava istih 100 zapisov traja po šest minut pri enaki vrednosti dela 30 € na uro in da prinese istih 90 sprejetih paketov. To bi stalo 300 € in porabilo 600 minut. Predpostavljamo, da ročna pot ne potrebuje dodatnega stroška programske opreme ali vzpostavitve; če ga potrebuje, ga dodajte. Primer s pomočjo modela porabi 430 minut človeškega dela neposredno na nalogah in skupaj z razporejenimi stroški stane 259,20 €: razlika znaša 170 minut neposrednega dela in 40,80 €. To sta posledici predpostavljenih enakih izidov, ne opažena prihranka. Kakovost in trajanje ročnega dela je treba izmeriti, preden lahko ta primerjava podpre uvedbo.

Pred dodajanjem izvajalcev poiščite ozko grlo

Predpostavimo dve sočasni izvajalni mesti v oblaku in štiri minute zasedenosti mesta na poskus modela. Obdelava 120 poskusov ob popolni razporeditvi med mesti potrebuje najmanj 240 minut. Lokalna predobdelava na eni napravi traja 150 minut. Človeško delo neposredno na nalogah skupaj traja 430 minut. Te stopnje se lahko delno prekrivajo, zato seštevanje vseh treh trajanj ne bi opisalo pretečenega časa celotnega toka. Tudi največje od teh trajanj ne zagotavlja časa dokončanja: pomembni so še odvisnosti, čakanje in zadnji pregled.

Koordinator ima za to tedensko delo na voljo osem ur oziroma 480 minut. Izmišljena serija pusti 50 minut rezerve človeškega časa. Dodatna izvajalna mesta v oblaku ne bi ustvarila več časa za pregled. Morda bi le prestavila čakalno vrsto od izvajalcev modela h koordinatorju. Če prispe 150 nalog, delavnica ne sme obljubiti vseh zgolj na podlagi ocene strojne zmogljivosti; delo nad obsegom sprejete vhodne serije mora izrecno odložiti, zanj zagotoviti ljudi ali ga zavrniti.

Dodatno mejo postavljajo omejitve ponudnika. Dokumentacija OpenAI ločuje omejitve hitrosti zahtev in žetonov, opozarja, da je mogoče doseči eno ali drugo, ter navaja, da se razlikujejo glede na model.[2] Naš izračun z dvema mestoma je le spodnja časovna meja razporejanja, ne pravica do ponudnikove zmogljivosti. Dejanski zagon mora beležiti hitrost porabe žetonov, čakanje na storitev, zamike pred ponovnimi poskusi in razpoložljivost pregledovalca ter tudi čas od začetka do konca.

Spremljajte starost najstarejše še odprte naloge in število nalog, ki čakajo na pregled. Končne zavrnitve in neuspehe ohranite v stroškovnem poročilu serije, vendar teh zaključenih nalog ne prištevajte k čakajočemu delu. Čakalna vrsta, ki raste iz tedna v teden, kopiči obveznost, tudi če vsak izvajalec poroča o visoki izkoriščenosti. Nekaj pregledovalčeve zmogljivosti rezervirajte za neobičajne primere in prekinitve, namesto da bi razporejali do teoretičnega maksimuma. Ko rok za sprejemno nalogo poteče, koordinator odloči, ali se delo nadaljuje; izvajalec ne sme tiho prestaviti datuma.

Pravilo ponovnih poskusov je del proračuna

Konkretna napaka lahko razveljavi izračune cene in zmogljivosti. Predpostavimo, da aplikacija dovoli en ponovni poskus, njen odjemalec API pa zahteve ponavlja tudi sam. Vidna čakalna vrsta še vedno poroča o dveh poskusih, do ponudnika pa prispe več zahtev. Stroški in pretečeni čas se ne ujemajo več z modelom serije. Trenutne smernice OpenAI izrecno pravijo, da je treba ponovne poskuse razvojnega kompleta SDK izključiti ali jih upoštevati, kadar jih upravlja aplikacija, da ugnezdene zanke ne pomnožijo zahtev.**[2]**

Predlagano pravilo delavnice dovoljuje en dodatni poskus modela na sprejeto vhodno nalogo, štet na meji dejansko odposlane zahteve. Prehodne napake storitve lahko povzročijo odlog znotraj roka. Manjkajoči izvorni podatki, zavrnjen dostop in vsebinska zavrnitev ne sprožijo slepih ponovnih poskusov. Google SRE kot

nadzor nad preobremenitvijo opisuje omejevanje ponovnih poskusov na zahtevo in na odjemalca; njegove številčne nastavitve veljajo za tam opisani sistem, ne za to delavnico.[1]

Pred začetkom rezervirajte sredstva za dovoljeni maksimum: 200 poskusov po 0,06 € potrebuje 12 € razpoložljivega proračuna za API. Predpostavljenih 120 dejanskih poskusov stane 7,20 €; neporabljenih 4,80 € sprostite šele po uskladitvi vseh še odprtih poskusov. Ta rezervacija za API ni dodatna stroškovna postavka. Neznan izid zahteve ostaja nerazrešena stroškovna izpostavljenost. Tudi okrevanje po omejitvi hitrosti mora upoštevati odgovor ponudnika: OpenAI dokumentira čakanje vsaj toliko časa, kot določa veljavni **Retry-After**, ter omejitev števila in skupnega trajanja ponovnih poskusov.[2] Kadar bi takšno čakanje preseglo rok naloge, jo mora izvajalec odložiti, ne pa »poskusiti še bolj zavzeto« s preostalim proračunom.

Spreminjajte po eno predpostavko

Analiza občutljivosti preveri, ali zmerna sprememba obrne odločitev. Naslednje možnosti spreminjajo le navedeno spremenljivko. Njihovi izidi so računski scenariji, ne napovedi. Denarni rezultati na sprejeti paket so po deljenju zaokroženi na cente.

Drugačna predpostavka

Osnovni primer

Strošek serije	259,20 €
Sprejeti paketi	90
Strošek na sprejeti paket	2,88 €

Cene API-ja se podvojijo; vse drugo ostane enako

Strošek serije	266,40 €
Sprejeti paketi	90
Strošek na sprejeti paket	2,96 €

Vsak začetni pregled traja minuto dlje

Strošek serije	309,20 €
Sprejeti paketi	90
Strošek na sprejeti paket	3,44 €

Enaka poraba, a le 75 sprejetih paketov

Strošek serije	259,20 €
Sprejeti paketi	75
Strošek na sprejeti paket	3,46 €

Polovični obseg in sorazmerno spremenljivo delo; razporejeni stalni stroški ostanejo enaki

Strošek serije	147,10 €
Sprejeti paketi	45
Strošek na sprejeti paket	3,27 €

Vrstica s polovičnim obsegom prepolovi spremenljive stroške in sprejeto delo, ohrani pa 35 € razporejenih stalnih stroškov. To je izračun pričakovane sestave nalog, ne dobesedna trditev, da je mogoče opraviti polovico določenega urejenega paketa. Vrstica s 75 sprejetimi paketi ohrani enako porabo, da loči učinek imenovalca; ne trdi, kateri potek neuspehov bi povzročil tak izid.

Pri teh predpostavkah čas pregleda in sprejem vplivata bolj kot podvojitev cen API-ja. To ne določa splošnega vrstnega reda za druge vrste dela. Raziskovalna naloga z dolgim kontekstom, drag podatkovni vir ali zahtevna lokalna obdelava bi lahko imeli drugačno stroškovno sestavo. Tabelo uporabite za odločitev, kaj izmeriti naslednje, ne za utemeljevanje izbire najcenejšega modela brez primerjave kakovosti.

Vnaprej določite pogoje za ustavitev

Izmišljeni lastnik pred preizkusom določi zgornjo mejo 3,20 € na sprejeti paket, zahteva vsaj 90 sprejetih nalog iz te serije in čas človeškega dela pri izvajanju omeji na 480 minut. To so vodstvene odločitve, ne splošni standardi. Osnovni primer na papirju izpolni vse tri. Upraviči omejen merilni preizkus, ne samodejne širitve.

Če pregled dosledno potrebuje dodatno minuto iz tabele občutljivosti, strošek na enoto naraste na 3,44 €, človeško delo pa na 530 minut. Obe omejitvi sta preseženi. Odločitev je začasno ustaviti nove serije, ohraniti vidnost nerazrešenih nalog in pred nadaljevanjem raziskati obremenitev pri pregledu. Že nastali strošek vzpostavitve se ne povrne z izvajanjem še več slabega dela. Spremenjeni delovni tok potrebuje novo primerjavo z ročnim izhodiščem in izrecno proračunsko odločitev.

Ta primer delavnice nazadnje ne predpostavlja nobenih prihodkov od prodaje. Če podjetje prodaja sprejete pakete, je prihodek tisto, kar kupci dejansko plačajo; prispevek za kritje nastane po odštetju spremenljivih stroškov izvedbe, preostale stalne stroške pa je še vedno treba pokriti. Ocena tehnične zmogljivosti ne dokazuje povpraševanja strank ali dobičkonosne cene. Evidenca delovanja naj pokaže sprejeto delo, neuspehe, človeški čas, strošek in razlog za nadaljevanje ali ustavitev, ne da bi dejavnost spreminjala v izmišljeni prihodek.

Viri

Številke virov veljajo za ta vodnik. Povezave vodijo do izvirnih dokumentov. Pred uporabo preverite besedilo, pogoje uporabe in razpoložljivost storitev.

[1] **Handling Overload — Site Reliability Engineering**
https://sre.google/sre-book/handling-overload/

[2] **Rate limits — OpenAI API**
https://developers.openai.com/api/docs/guides/rate-limits

[3] **Amazon SQS at-least-once delivery**
https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/standard-queues-at-least-once-delivery.html