KOFETKE.

Prenos uporabnega delovnega toka med ponudniki

Kako se izogniti ponovni gradnji vsega, ko se spremeni model ali cena?



PRENOSLJIVOST DELOVNIH TOKOV

Prenosljiva pogodba za vhode in izhode ter preskus selitve.

SLOVENSKA IZDAJA

september 2026
Praktični vodniki o uporabi AI

Viri preverjeni 22. septembra 2026.

Uporabnega delovnega toka z AI ne bi smelo biti treba graditi znova zato, ker je model ukinjen, se spremeni cena, regija postane neustrezna ali ponudnik doda funkcijo, ki je pomembna za delo. To ne pomeni, da je mogoče vsak model zamenjati z eno spremenljivko okolja. Pomeni, da ima delovni tok del, ki pripada podjetju, in manjši del, ki pripada določenemu ponudniku.

Del v lasti podjetja opiše opravilo: odobreni vhod, pričakovani izhod, mejo pregleda, zapise, ki jih je treba hraniti, ter pogoje, ki pomenijo uspeh ali napako. Del, vezan na ponudnika, to opravilo prevede v obliko zahtevka, orodja, preverjanje pristnosti, omejitve in obliko odgovora posamezne storitve. Če ti plasti ostaneta ločeni, je prehod mogoč. Rezultat pa zaradi tega ni enak.

API, označen kot združljiv, je uporaben za preizkus nove storitve. Anthropicove opombe ob izdaji na primer opisujejo končno točko, združljivo z OpenAI, za preizkušanje osnovne funkcionalnosti chat-completions ob zamenjavi ključa API, osnovnega URL-ja in imena modela.**[3]** Ta izjava je namenoma ozka. Zahtevek lahko doseže drugo končno točko, medtem ko se izbira orodij, vedenje pri zavrnitvi, funkcije JSON, omejitve hitrosti, štetje žetonov, varnostne kontrole in kakovost izhoda še vedno razlikujejo. Združljivost na ravni prenosa ni vedenjska enakovrednost.

Začnite s pogodbo, ki je v lasti delovnega toka

Prenosljiva pogodba je majhen dokument z različico, ki opiše, kaj vstopa v delovni tok in kaj iz njega izstopa. Ne sme biti kopija celotnega zahtevka enega ponudnika. Povedati mora, kaj mora vedeti preostanek podjetja.

Poglejmo izmišljen primer. Maja vodi majhno veleprodajo. Njen delovni tok prebere e-pošto dobavitelja in pripravi postavko za pregled. Nikoli ne knjiži računa in ne spreminja računovodskega sistema. Pogodba ima štiri plasti:

Plast	Za kaj je odgovorna	Primer
Opravilo	Namen in pooblastilo	Izlušči možne podatke računa za človeškega pregledovalca.
Vhod	Stabilni poslovni podatki	Besedilo sporočila, odobrene priloge, ID vira in jezik.
Izhod	Zahtevani zapis odločitve	Možna polja, negotovosti, navedbe izvornih strani in **needs_review**.
Politika	Pravila zunaj modela	Brez knjiženja, brez odgovora dobavitelju, brez izmišljene davčne kode in brez ugibanja manjkajočih polj.

Izhod je mogoče izraziti z majhno shemo. To je ponazoritvena pogodba, ne zahtevek ponudniku:

```
{
  "contract_version": "invoice-review/v1",
  "source_id": "mail-2026-0917-044",
  "currency": "EUR",
  "total": null,
  "candidate_fields": [],
  "uncertainties": [],
  "needs_review": true
}
```

Pogodba razlikuje med manjkajočo vrednostjo in ugibano vrednostjo. **null** pomeni, da delovni tok zneska ni ugotovil. Prazno polje **candidate_fields** pomeni, da ni našel utemeljenega kandidata. **needs_review** ni ocena zaupanja, preoblečena v dovoljenje za knjiženje. Te razlike se med ponudniki prenesejo bolje kot dolg poziv modelu, naj bo »previden«.

Izhod, omejen s shemo, lahko zmanjša eno vrsto napake pri integraciji. OpenAI opisuje Structured Outputs kot izhode, ki se držijo podane sheme JSON, medtem ko starejši način JSON zagotavlja le veljaven JSON, ne pa skladnosti s shemo.[5] To je lahko uporabno na meji med modelom in aplikacijsko kodo. Ne dokazuje pa, da so znesek, datum ali ime dobavitelja v poljih pravilni. Strukturno veljaven napačen račun je še vedno napačen.

Tudi preverjanje domene naj ostane zunaj modela. Aplikacija lahko preveri, ali je valuta dovoljena, ali se datum razčleni, ali se izračunana vsota postavk ujema z navedenim skupnim zneskom znotraj določenega pravila za zaokroževanje in ali vsak navedeni

del strani obstaja v odobreni prilogi. Ti preskusi so prenosljivi, ker preverjajo lastno pogodbo delovnega toka, ne proze posameznega ponudnika.

Česa prilagoditveni sloj sme in česa ne sme skriti

Prilagoditveni sloj preslika prenosljivo pogodbo v ponudnika. Lahko izbere model, prevede sistemska navodila, zahteva strukturirani izhod, doda opise orodij, nastavi časovno omejitev, normalizira odgovor in zabeleži metapodatke ponudnika. Ne sme pa tiho spremeniti poslovnega pravila, ker ima novi ponudnik drugačno priročno funkcijo.

Uporaben vmesnik prilagoditvenega sloja ima izrecne operacije, kot so **prepare_request**, **run**, **normalise_response** in **record_usage**. Normalizirani odgovor vsebuje izhod pogodbe, ime ponudnika, identifikator modela, čas zahtevka, trajanje, podatke o žetonih ali uporabi, če so na voljo, klice orodij, kategorijo zavrnitve ali napake ter mesto surovega odgovora, kadar je njegovo hranjenje odobreno. Polja ponudnika hranite ob prenosljivem zapisu, ne v njegovem pomenu.

Ta ločitev pomaga tudi pri identitetah modelov. Imena ponudnikov pogosto vključujejo vzdevke, datirane posnetke in različne identifikatorje na različnih gostujočih platformah. Anthropicov pregled modelov napotuje na ločeno dokumentacijo za ID-je modelov, vzdevke in posnetke ter navaja, da lahko API za modele vrne omejitve in podatke o zmogljivostih.[4] Zapis o selitvi

naj zato za vsak zagon vrednotenja navede natančni identifikator modela, različico prilagoditvenega sloja in datum. »Preskusili smo hitri model« ni ponovljivo.

Ne silite vseh funkcij skozi vmesnik najmanjšega skupnega imenovalca. Ponudnik lahko ponuja posebno orodje za iskanje, predpomnilnik, obliko slikovnega vhoda, varnostno kontrolo ali paketno zmogljivost, ki je drugi ne ponuja. Takšno funkcijo postavite za imenovano izbirno zmogljivost. Če je delovni tok od nje odvisen, odvisnost navedite in preskusite nadomestno pot. Če varne nadomestne poti ni, delovni tok ni prenosljiv k ponudniku brez te zmogljivosti. Skrivanje tega dejstva ustvari selitev, ki je na diagramu čista, v resničnem primeru pa odpove.

Vedenje potrebuje preskuse, ne optimizma

Preskus pogodbe vpraša, ali aplikacija lahko izmenja zahtevane podatke. Vedenjski regresijski preskus vpraša, ali premaknjen delovni tok še vedno opravlja delo varno in uporabno. Potrebna sta oba.

OpenAIjev vodnik za vrednotenja pravi, da so vrednotenja posebej pomembna pri nadgradnji ali preizkušanju novih modelov, ter opisuje zanko opredelitve opravila, zagona preskusnih vhodov, analize rezultatov in izboljšanja poziva.[1] Anthropicov vodnik za vrednotenje dodaja, da večina primerov uporabe potrebuje več meril uspeha in med proračunskimi vprašanji izrecno omenja ceno, velikost modela in pogostost uporabe.[2] Ta vira opisujeta lastni platformi, vendar načelo ostane: sprejem selitve ima več razsežnosti.

Za Majin delovni tok pred prehodom zgradite zamrznjen nabor preskusov. Vsebovati mora odobrene, razosebljene primere, ki zajemajo običajne račune, manjkajoče skupne zneske, podvojene priloge, več valut, sporočilo z zahtevo, naj sistem prezre navodila, sporočilo brez računa, zavajajočo številko v e-poštnem podpisu, ime dobavitelja v nelatinični pisavi in namerno neberljivo prilogo. Za vsak primer ohranite pričakovani izid, ki ga je pregledal človek. Nabora ne sestavite samo iz lahkih e-poštnih sporočil, zaradi katerih je stari delovni tok videti uspešen.

Uporabite več družin preskusov:

Družina preskusov	Vprašanje	Primer napake
Pogodba	Ali se izhod razčleni in vsebuje vsako zahtevano polje?	Manjka zahtevani ključ **needs_review**.
Celovitost podatkov	Ali izhod ohrani ID-je virov in razlikuje neznano od ugibanega?	Manjkajoči skupni znesek postane **0,00**.
Vedenje opravila	Ali prepozna utemeljene kandidate in označi dvoumnost?	Telefonsko številko izbere kot znesek računa.
Varnost in orodja	Ali zavrne neodobrena dejanja in se upre sovražnemu besedilu priloge?	Sledi vdelani zahtevi, naj dobavitelju pošlje e-pošto.

Družina preskusov	Vprašanje	Primer napake
Ravnanje ob napaki	Ali časovno omejitev, zavrnitev ali napačen rezultat orodja razvrsti brez prikrivanja?	Napaka ponudnika postane dokončana postavka pregleda.
Operativne omejitve	Ali ostane znotraj omejitev časa, stroška in sočasnosti?	Primeri uspejo, vendar preseže dovoljeni strošek zagona.

Iste zamrznjene primere zaženite skozi sedanji prilagoditveni sloj in kandidata. Primerjajte več kot eno samo povprečno oceno. Poglejte vsako varnostno napako, vsak manjkajoč sklic na vir, vsak popravek pregledovalca, porazdelitev zakasnitev, strošek na sprejeto postavko in kategorije napak. Razlike so lahko pričakovane; nepojasnjene razlike so razlog za preskus.

Pri ustvarjalnem opravilu pričakovani rezultat morda ni en sam ustrezen stavek. Preskus lahko preveri, ali so prisotna vsa zahtevana dejstva, ali se ne pojavi nobeno nepodprto dejstvo, ali ton ostane znotraj določene meje in ali pregledovalec rezultat sprejme. Vzorec človeškega pregleda ostane potreben, kadar je opravilo odvisno od pomena, presoje ali zaupanja bralca. Model, ki ocenjuje drug model, lahko pomaga pri razvrščanju, ne sme pa biti edina zapora pri varnostni ali poslovni odločitvi.

Selitev, ki jo je mogoče obrniti

Majina selitev ima pet stopenj. Najprej zamrzne **invoice-review/v1**, nabor preskusov, različico sedanjega prilagoditvenega sloja in sedanja pravila sprejema. Nato zgradi prilagoditveni sloj kandidata, ne da bi spremenila politiko delovnega toka. Tretjič oba prilagoditvena sloja zažene na zamrznjenih primerih ter zabeleži natančne identitete modelov in stroške. Četrtič popravi samo upravičene razlike v prilagoditvenem sloju ali pozivu, znova zažene preskuse in dokumentira vsako sprejeto spremembo vedenja. Petič majhen paket samo za pregled usmeri h kandidatu, medtem ko stara pot ostane na voljo za povrnitev.

Paket samo za pregled ni dokaz uspeha v produkciji. Je način, kako odkriti težave z resničnimi trenutnimi vhodi, medtem ko človek še vedno preveri vsak rezultat in noben izhod ne more knjižiti v računovodski sistem. Prehod mora imeti naveden pogoj za povrnitev: na primer vsak poskus neodobrenega dejanja, vsaka kršitev pogodbe po normalizaciji ali regresija varnostnega preskusa novo delo vrne k prejšnjemu odobrenemu prilagoditvenemu sloju. Zapišite razlog. Nabora preskusov hkrati ne spreminjajte samo zato, da bi kandidat uspešno opravil preskus.

Pogodba z različico prepreči tudi pogosto past. Recimo, da novi ponudnik ne more vrniti navedenega odseka vira v isti obliki. Morda je potreben **invoice-review/v2**, pri katerem izhod namesto tega poda stran priloge in oznako polja. To je sprememba izdelka, ne tiha sprememba prilagoditvenega sloja. Stari in novi različici

naj obstajata hkrati, dokler niso preskušeni odvisni pregledovalci in sistemi. Prenosljivi delovni tok se sme razvijati; razvoja ne sme prikriti kot zamenjavo ponudnika.

Primer napake: veljaven JSON, napačno delo

Predstavljajte si, da prilagoditveni sloj kandidata za e-pošto, ki vsebuje telefonsko številko dobavitelja, datum dostave in nobenega skupnega zneska, vrne ta popolnoma veljavni zapis:

```
{
  "contract_version": "invoice-review/v1",
  "source_id": "mail-2026-0917-044",
  "currency": "EUR",
  "total": 421.50,
  "candidate_fields": ["421.50"],
  "uncertainties": [],
  "needs_review": false
}
```

Struktura uspe. Delovni tok odpove. Število je prišlo iz telefonske interne številke ali iz nepovezane stare ponudbe, sistem pa je potrdil rezultat, ki ga vir ne podpira. Preskus sheme ne ujame niti napačnega dokaza niti neustrezne vrednosti **needs_review**. Zamrznjeni primer potrebuje pravilo dokaza: vsak kandidat za znesek mora navesti mesto v viru, noben znesek pa ne sme biti sprejet, kadar vir nima skupnega zneska računa. Pričakovani izid, ki ga je pregledal človek, je **total: null**, negotovost z razlago in **needs_review: true**.

Ta primer pojasni, zakaj združljivi zahtevki niso dovolj. Izhodi modelov so verjetnostni, izvedbe ponudnikov se spreminjajo, vzdevki modelov pa lahko kažejo na nove različice. Preskus selitve mora zato zaščititi poslovno vedenje, ki je bilo pomembno pred prehodom, ne le skladnje zahtevka.

Vodite zapis za izhod

Za vsako pot ponudnika ohranite kratek zapis za prehod k drugemu ponudniku: različico pogodbe, različico prilagoditvenega sloja, ID modela, zahtevane zmogljivosti, različico nabora preskusov, sprejete pragove, datum zadnjega pregleda, predpostavke o stroških, omejitve lokacije podatkov, če so pomembne, in cilj povrnitve. Pozive in opise orodij vodite pod enakim nadzorom sprememb kot kodo. Sprememba poziva lahko spremeni vedenje toliko kot sprememba modela.

Zapis omogoča tudi pošteno primerjavo. Ponudnik z nižjo ceno žetona lahko stane več na sprejeto postavko, če ustvari več dela za pregledovalca. Ponudnik z zmogljivim orodjem je lahko neustrezen, če se njegovo zahtevano ravnanje s podatki razlikuje od odobrene meje. Ponudnik, ki uspe na vseh strukturnih preskusih, lahko še vedno odpove na jezikovnih ali varnostnih preskusih. Univerzalnega zmagovalca ni; obstaja dokumentirana izbira za določeno opravilo.

Izognite se preskusu selitve, ki po naključju spremeni pet stvari hkrati. Če pot kandidata uporablja nov model, prepisan poziv, drugačna orodja, nov razčlenjevalnik dokumentov in spremenjen zaslon za pregledovalca, boljšega rezultata ni mogoče pripisati eni sami spremembi. Med primerjavo prilagoditvenih slojev naj

pogodba, testni primeri in pravila točkovanja ostanejo nespremenjeni. Če nova pot potrebuje drugo zmogljivost, ustvarite ločen preskus, ki navede zmogljivost in njen pričakovani učinek. Tako je poznejšo regresijo mogoče diagnosticirati, namesto da ostane skrivnostna.

Tudi nabor preskusov potrebuje skrbništvo. Vsakemu primeru dodelite ID, različico vira, razlog za vključitev in pričakovano stanje. Težke protiprimere ohranite tudi potem, ko ponudnik na njih odpove; izbris neprijetnega primera spremeni znano slabost v nevidno. Sveže primere dodajte le z zabeleženo spremembo, nato pa znova zaženite obe poti, kjer je ta primerjava še pomembna. Namen ni zamrzniti napredka. Namen je ohraniti pomen uspešnega preskusa skozi čas.

Prenosljivost ni trditev, da so storitve zamenljive. Je praksa lastništva nad opredelitvijo opravila, ohranjanja preskusljive pogodbe o izhodu, ločevanja prilagoditvenih slojev ponudnikov in dokazovanja naslednje poti z resničnim vedenjem, preden dobi pooblastilo. To zadošča, da sprememba modela ali cene postane nadzorovano vzdrževalno opravilo namesto ponovne gradnje.

Viri

Številke virov veljajo za ta vodnik. Povezave vodijo do izvirnih dokumentov. Pogoji uporabe in razpoložljivost storitev se lahko spremenijo.

[1] **Working with evals | OpenAI API**
https://platform.openai.com/docs/guides/evals

[2] **Define success criteria and build evaluations | Anthropic**
https://platform.claude.com/docs/en/test-and-evaluate/develop-tests

[3] **Claude Platform release notes | Anthropic**
https://platform.claude.com/docs/en/release-notes/overview

[4] **Models overview | Anthropic**
https://platform.claude.com/docs/en/about-claude/models/overview

[5] **Structured model outputs | OpenAI API**
https://platform.openai.com/docs/guides/structured-outputs