KOFETKE.

Vrednotenje pomočnika na lastnem delu

Kako primerjate sisteme, ne da bi zaupali lestvici?



VREDNOTENJE POMOČNIKOV

Preizkusni niz za določeno nalogo z napakami, časom pregleda in sprejetimi izidi.

SLOVENSKA IZDAJA

september 2026
Praktični vodniki o uporabi AI

Viri preverjeni 22. septembra 2026.

Javna lestvica je lahko zanimiva, vendar fotografu v Sloveniji ne pove, ali bo pomočnik iz njegovega gradiva pripravil uporabne zapiske za fotografiranje. Majhni računovodski pisarni tudi ne pove, ali bo pomočnik račun pravilno usmeril v pregled, ne da bi izgubil davčno številko. Lestvica poda rezultat za tujo nalogo, tuje podatke, tuja pravila in tuje ocenjevanje. Vaša odločitev je bliže zaposlitvi začasnega sodelavca: ali lahko ta določena ureditev pripravi sprejemljivo delo za naloge, ki so tu pomembne?

Za to potrebujete majhno vrednotenje, pogosto imenovano evalvacija. Evalvacija je ponovljiv preizkus z določenimi vnosi, pričakovanim rezultatom ali pravilom pregleda in zabeleženim izidom. Ni predstavitev, pri kateri pomočnik po naključju poda všečen odgovor. Preizkus mora omogočati primerjavo možnosti. Smernice OpenAI razlikujejo med preizkusi za določeno nalogo in splošnimi merili ter priporočajo, da preizkusi odražajo porazdelitve iz resnične rabe. Njihov ločen vodič opisuje zadržani preizkusni niz kot del podatkov, ki ostane nedotaknjen in lahko razkrije prilagajanje merilu.**[1][2]**

Članek ponuja metodo za omejen poslovni proces. Ne določa splošne ocene in ne obljublja, da bo en model ostal najboljši, ko ponudnik spremeni model, orodje, ceno ali pravila. Uporaben rezultat je zapis odločitve: katera konfiguracija je bila preizkušena, na katerih primerih, kaj je šlo narobe, koliko preverjanja je bilo potrebno in ali je bilo dokončano delo sprejeto.

Začnite z odločitvijo, ne z imenom modela

Izberite eno nalogo z vidnim koncem. »Pomoč pri administraciji« je preširoko. »Preberi e-pošto dobavitelja in pripravi osnutek zapisa za človeškega knjigovodjo« je dovolj ozko. Dobra prva naloga ima omejen vhod, znanega pregledovalca in izid, ki ga je mogoče sprejeti, popraviti, zavrniti ali predati v nadaljnjo obravnavo.

Pred preizkusom napišite kratko kartico naloge:

Polje	Primer za pomočnika	Zakaj je pomembno
Naloga	Pripravi kartico za pregled e-pošte dobavitelja	Prepreči nejasno primerjavo
Vhod	E-pošta in besedilo priloženega računa	Določa, kaj sistem sme uporabiti
Dovoljeno dejanje	Samo osnutek; brez knjiženja ali plačila	Loči pooblastilo od pomoči
Zahtevani izhod	Dobavitelj, znesek, valuta, rok, povezave do dokazov, neznanke	Pregled naredi konkreten
Pregledovalec	Imenovani knjigovodja	Določi lastnika sprejema
Pravilo sprejema	Vsa podprta polja so pravilna; neznana polja so označena; brez izmišljenega zneska	Uspeh določi pred ogledom rezultatov

Polje	Primer za pomočnika	Zakaj je pomembno
Pravilo prekinitve	Prekini ob nepodprtem finančnem podatku	Omeji ponavljanje napak, ki se jim je mogoče izogniti

Kartica naloge razkrije tudi dele, o katerih model ne more odločiti. Račun je lahko nečitljiv. Rok plačila morda manjka. Pravilno dejanje v teh primerih ni ugibanje, ampak jasno označena neznanka in predaja osebi. Preizkus mora priznati to zadržanost. Sicer je lahko zgovorno izmišljena vrednost videti boljša od poštene predaje.

Pričakovani izhod ni nujno en popoln odstavek. Pri strukturiranem delu je lahko kontrolni seznam, nabor polj, klic orodja ali končno stanje. Pri odprtem pisanju uporabite kratek rubrični opis s primeri sprejemljivega in nesprejemljivega dela. Smernice za vrednotenje OpenAI priporočajo združevanje meritev s človeško presojo ter umerjanje samodejnega točkovanja s povratnimi informacijami ljudi.[2] To je priporočilo za zasnovo, ne dokaz, da je določen samodejni ocenjevalnik zanesljiv za vašo nalogo.

Primere zgradite iz dela, ki ga že poznate

Najvrednejši primeri pogosto izhajajo iz dokončanega dela, popravljenih osnutkov in pogostih izjem. Pred uporabo odstranite ali zamenjajte osebne podatke. Ohranite lastnosti, zaradi katerih je bil primer težaven: sporočilo z dvema računoma, manjkajočo prilogo, večjezično opombo, datum, ki ga je mogoče razumeti na dva načina, ali dobavitelja, ki je spremenil ime.

Primere razdelite v tri mape.

```
znani primeri       -> za snovanje navodil
regresijski primeri -> stare napake; znova po spremembi
zadržani primeri    -> zaprti do dneva primerjave
```

Znani primeri pokažejo, kje delovni tok odpove. Regresijski primeri preprečujejo, da bi se odpravljena napaka neopazno vrnila. Zadržani primeri niso tajni preizkus, ki ga pripravi ponudnik. So navadni ustrezni primeri, ki jih oseba, ki spreminja navodila, orodja ali primere, med prilagajanjem ne pregleda. Če se rezultat izboljša na znanih primerih, ne pa na zadržanih, je razumna ugotovitev, da se je konfiguracija morda prilagodila preizkusu. To ni dokaz splošne kakovosti.**[1]**

Map ne urejajte umetno. Vključite običajno delo, nerodno delo in delo, ki bi ga pomočnik moral zavrniti ali predati. Preizkusni niz, sestavljen samo iz čistih računov, odgovori le na vprašanje, ali pomočnik obvlada čiste račune. V dejanskem procesu manjkajoče informacije in nasprotujoča navodila pogosto določajo obseg pregleda. Priročnik NIST AI RMF svetuje dokumentiranje preizkusnih nizov, meritev in orodij ter merjenje sistemov v razmerah, podobnih predvidenemu okolju uporabe.**[3]**

Majhen niz je uporaben, kadar je premišljeno izbran. Dvanajst primerov lahko odkrije ponavljajočo se napako. Dvanajst primerov ne more določiti tržne lestvice, zanesljivega odstotka za vse prihodnje delo ali varnostnega jamstva. Zapišite velikost vzorca in sestavo primerov, namesto da bi majhen rezultat olepševali z decimalnimi mesti.

Izvedite primerljive poskuse

Primerjava je poštena le, če vsaka možnost dobi isto kartico naloge, vhod, dostop do orodij, obliko izhoda in pravilo pregleda. Zapišite različico ali datum modela, nabor navodil, povezana orodja in vsak vir za iskanje. Če ena možnost dobi jasnejše navodilo ali dodaten dokument, to označite kot del preizkušene konfiguracije.

Če je sistem spremenljiv, uporabite več kot en zagon, vendar spremenljivosti ne spreminjajte v tekmovanje, ki nagrajuje največ poskusov. Praktičen način je, da število poskusov določite vnaprej, označite vsak poskus in pregledate vse. Za pomočnika, ki pripravi eno kartico za pregled, je lahko zapis takšen:

Primer

Dva računa v enem e-sporočilu

Konfiguracija	A
Sprejeto pri prvem pregledu?	Ne
Napaka ali negotovost	Združil je dva roka plačila
Minute pregleda	6
Končni izid	Popravljeno in sprejeto

Dva računa v enem e-sporočilu

Konfiguracija	B
Sprejeto pri prvem pregledu?	Da
Napaka ali negotovost	Ni najdene napake
Minute pregleda	3
Končni izid	Sprejeto

Manjkajoča priloga

Konfiguracija	A
Sprejeto pri prvem pregledu?	Da
Napaka ali negotovost	Označil je manjkajočo prilogo
Minute pregleda	2
Končni izid	Pravilno predano

Manjkajoča priloga

Konfiguracija	B
Sprejeto pri prvem pregledu?	Ne
Napaka ali negotovost	Izmislil si je znesek
Minute pregleda	5
Končni izid	Zavrnjeno

Izmišljene številke v tabeli kažejo obliko zapisa, ne rezultat raziskave. Nikoli jih ne uporabljajte kot obljubljen prihranek. Glavno vprašanje ni, ali je bil izhod videti uglajen. Pomembno je, ali je pregledovalec dokončano postavko sprejel po pravilih, določenih pred poskusom.

Spremljajte najmanj štiri stvari. Prvič, zapišite sprejete izide: število sprejetih brez pomembnega popravka, sprejetih po popravku, zavrnjenih in pravilno predanih primerov. Drugič, napake razvrstite po vrsti, na primer napačno izluščenje, nepodprta trditev, napačna izbira orodja, napaka oblikovanja ali spregledana predaja. Tretjič, zabeležite čas pregleda od odprtja izhoda do odločitve. Četrtič, ohranite skrbništvo nad dokazi: izvorni vhod ali zakonito redigirano referenco, izhod pomočnika, sled orodja, kadar je pomembna, odločitev pregledovalca in različico kartice naloge.

Čas pregleda je preprosto napačno razlagati. Ni samodejno prihranjen čas dela. Pregledovalec lahko porabi manj časa za preverjanje napačnega odgovora, ker je napaka prikrita, ali več časa za uporaben odgovor, ker je primer izjemno zapleten. Čas pregleda primerjajte skupaj s sprejemom in vrstami napak. Če se čas zmanjša, nepodprte trditve pa naraščajo, proces morda hitreje prenaša tveganje naprej.

Točkujte previdno in preberite primere

Preprosta stopnja izidov je lahko uporabna:

```
stopnja sprejetih izidov = sprejeti izidi / upravičeni
primeri
```

Če je na primer 9 od 15 upravičenih primerov sprejetih po določenem pravilu, je stopnja 9 / 15 = 0,60 oziroma 60 %. To opisuje ta zagon na teh 15 primerih. Ni napoved za vse delo. Imenovalec sme izključiti le primere, ki so bili pred pregledom določeni kot neupravičeni, na primer poškodovano datoteko, ki je nobena konfiguracija ni mogla odpreti. Ne odstranjujte neprijetnih neuspehov, potem ko jih vidite.

Drugi izračun lahko pokaže obseg pregleda:

```
mediana časa pregleda = srednja zabeležena vrednost časa
pregleda
```

Mediana je pogosto koristnejša od povprečja, kadar en zahteven primer traja veliko dlje od drugih. Prikažite tudi posamezne čase ali majhno porazdelitev, saj lahko en povzetek skrije, da nekaj primerov povzroča večino dela.

Samodejne kontrole so dragocene tam, kjer je dejstvo mehansko: zahtevana polja so prisotna, datum je v zahtevani obliki, shema JSON je veljavna ali argument orodja ustreza znanemu identifikatorju. Pregledovalca uporabite za trditve, ki so odvisne od razlage vira, poslovne politike ali resnične uporabnosti odgovora. Vodič za vrednotenje navaja deterministične ocenjevalnike za strojno preverljive lastnosti in opozarja, da je treba ocenjevalnike, ki temeljijo na modelu, pred razširitvijo preveriti s človeškimi oznakami.**[1][2]**

Preberite vsako zavrnitev in vzorec sprejetih primerov. Vzorec napak lahko pokaže spremembo zasnove, ki je skupni rezultat ne razkrije: pojasnite polje, odstranite nevarno orodje, dodajte referenčni dokument, vsilite neznano vrednost ali nalogo ohranite pri človeku. Kadar je mogoče, spremenite eno pomembno spremenljivko naenkrat, nato znova zaženite regresijske primere. Zadržani niz pustite nedotaknjen do smiselne primerjave.

Izmišljen primer: studijski predal za račune

Maja vodi izmišljeni studio za produktno fotografijo. Enkrat tedensko prejme e-pošto dobaviteljev in želi, da pomočnik pripravi kartice za pregled, ne pa da odobri stroške ali ustvarja računovodske vpise. Izbere 18 dokončanih, redigiranih primerov: 10 običajnih računov, 4 e-pošte z manjkajočimi podatki, 2 podvojena računa in 2 sporočili, ki sploh nista računa.

Maja in njen knjigovodja pred preizkusom določita uspeh. Kartica mora vsak izluščen podatek povezati s posredovanim besedilom, ohraniti izvirno valuto, označiti negotovost in nikoli ne sme v prazno polje vpisati številke. Knjigovodja lahko kartico sprejme, popravi, zavrne ali preda naprej. Opozorilo o dvojniku šteje kot koristno samo, če ga podpira izvirna referenca računa.

Konfiguracija A je en pomočnik s stalnim navodilom in orodjem za iskanje po dokumentih. Konfiguracija B doda primer odgovora ob manjkajoči prilogi. Obe dobita isti paket vhodov in nimata dostopa do računovodskega sistema. Maja prvih 12 primerov označi kot primere za zasnovo, šest pa zapre kot zadržane. Zapiše natančna navodila in čas za vsak pregled.

Predpostavimo, da B pripravi čistejše odgovore na 12 primerih za zasnovo, vendar si v enem zadržanem primeru z manjkajočo prilogo izmisli znesek za plačilo. Pravilen sklep ni, da je B boljši, ker je bil njegov prvi rezultat višji. Izmišljeni znesek je pomembna napaka po pravilu sprejema. Maja se lahko odloči, da obdrži A, spremeni navodilo B ali kartico oblikuje tako, da je polja za znesek ob manjkajoči prilogi nemogoče izpolniti. Vrednotenje je opravilo svojo nalogo, če je to odločitev naredilo vidno.

Ohranite zapis, podprt z dokazi

Ohranite kratek zapis vrednotenja z datumom, kartico naloge, identifikatorji primerov, obravnavo zasebnosti, identiteto konfiguracije, surovimi ali redigiranimi izhodi, odločitvami pregledovalca, izračuni in nerešenimi vprašanji. NIST opisuje dokumentiranje kot del ponovljivega merjenja in navaja, da lahko k rednemu ocenjevanju prispevajo neodvisni presojevalci ali notranji strokovnjaki, ki niso bili neposredni razvijalci.[3] Za majhno podjetje lahko neodvisnost pomeni preprosto to, da oseba, ki je napisala navodila, svojega dela ne označi sama kot sprejetega.

Ta zapis pomočnika ne potrjuje. Omogoča preverjanje omejene trditve: v teh pogojih, na teh primerih je ta konfiguracija ustvarila te izide. Preizkus ponovite, ko se naloga spremeni, ponudnik spremeni pomemben del sistema ali dejansko delo pokaže novo napako. Namen ni lov na imenitno merilo. Namen je vedeti, ali si določen pomočnik zasluži omejeno mesto v procesu, ki ga še vedno obvladujete.

Viri

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

[1] Realtime Eval Guide
https://developers.openai.com/cookbook/examples/realtime_eval_guide

[2] Evaluation best practices | OpenAI API
https://developers.openai.com/api/docs/guides/evaluation-best-practices

[3] NIST AI RMF Playbook: Measure
https://airc.nist.gov/airmf-resources/playbook/measure/