

3 promptai, kad AI neprigalvotų

AI klaidingą atsakymą paduoda taip pat užtikrintai kaip teisingą. Kodėl taip nutinka paprastais žodžiais ir trys promptai, kurie pagauna prasimanymą, kol dar juo nepasinaudojai.

PROMPTAI

7 MIN SKAITYMO

2026-09-16

Paklausi AI, o jis atsako švariai ir užtikrintai: su data, skaičiumi, gal net šaltiniu, tai ir panaudoji. O paskui paaiškėja, kad to niekada nebuvo. Tai vadinama haliucinacija: AI sako netiesą visiškai ramiu tonu, o pavojingiausia, kad klaidingas atsakymas atrodo lygiai taip pat kaip teisingas, be jokio įspėjamojo ženklų.

01

Kodėl AI prigalvoja

AI neieško atsakymo faktų knygoje. Iš esmės jis spėja labiausiai tikėtiną kitą žodį, panašiai kaip žodžių siūlymas telefone, tik daug galingesnis. Jis labai gerai moka sudėti sakinį, kuris skamba teisingai, o tai visai kitas įgūdis nei žinoti, kas iš tiesų tiesa.

Yra ir antra priežastis. Šie modeliai mokomi būti kuo naudingesni, o noras padėti stumia duoti atsakymą, o ne pasakyti „nesu tikras“. Kai modelis tema bent iš dalies atpažįsta, jis linkęs atsakyti, net jei faktų iš tikrųjų neturi. Tai skamba užtikrintai, nes dalis jo mano, kad žino.

@APIX.STUDIJA · APIXSTUDIJA.LI/GIDAI/3-PROMPTAI-KAD-AI-NEPRIGALVOTU/ · NEMOKAMAS GIDAS,

DALINKIS SU NUORODA

Kada tai nutinka dažniausiai. Su tiksliais skaičiais, datomis, pavardėmis ir šaltiniais, ir su labai siauromis arba labai naujomis temomis, kurių modelis mokydamasis matė mažai. Kuo siauresnis ar naujesnis klausimas (VMI deklaravimo terminas, Sodros įmokos dydis, teismo sprendimo numeris, praėjusios savaitės naujiena), tuo labiau verta atsakymą patikrinti.

02

Promptas 1: tegul pagrindžia kiekvieną teiginį

Pirmas vaistas paprastas: neleisk teiginiams kabėti ore. Priversk AI prie kiekvieno dalyko pridėti tikrą šaltinį, o tada nueik ir patikrink, ar šaltinis tikrai tai sako. Daug prasimanytų atsakymų subyra tą pačią sekundę, kai paklausi „iš kur tai“.

Naudok iškart po atsakymo, kuriame yra faktų, skaičių ar šaltinių. Įklijuok kaip kitą žinutę tame pačiame pokalbyje. Kad AI galėtų pats atidaryti nuorodas, pokalbyje turi veikti interneto paieška [patvirtinti: ar interneto paieška įjungta nemokamame Claude plane iš Lietuvos].

Prie kiekvieno teiginio, kurį ką tik parašei, pridėk tikslų šaltinį. Tada atidaryk kiekvieną šaltinį ir patikrink, ar jame tikrai parašyta tai, ką teigai. Ko negali patikrinti, pašalink arba aiškiai pažymėk.

Galutinį atsakymą parodyk dviem dalimis:

1. Teiginiai, kuriuos patvirtina tikras šaltinis, su šaltiniu prie kiekvieno.
2. Teiginiai, kurių patikrinti negalėjai, aiškiai pažymėti, kad žinočiau jais dar nepasitikėti.

Jeigu duoda šaltinį, bet nuoroda neatsidaro arba joje parašyta kas kita, tai raudona vėliava. Prasimanyti šaltiniai dažni, o tinka tik toks, kurį gali atsidaryti ir perskaityti pats.

03

Promptas 2: leisk jam nežinoti

Modelis mokomas padėti, ir tas spaudimas stumia jį duoti bent ką nors, o ne pripažinti spragą. Tai nuimk tą spaudimą. Pasakyk tiesiai, kad „nežinau“ yra visiškai geras atsakymas ir kad tau geriau spraga nei užtikrinta klaida.

Skamba per paprastai, bet veikia. Kai nebereikia visada turėti atsakymo, modelis daug noriau sustoja ir pasako, kad nėra tikras, užuot užpildes

spragą spėjimu.

Naudok pokalbio pradžioje, kai tikslumas svarbiau už greitį. Įklijuok prieš tikrąjį klausimą arba kaip taisyklę pačiame viršuje.

Prieš atsakydamas, viena taisyklė visam šitam pokalbiui:

Jei nesi tikras arba neturi faktų, sakyk „nežinau“, o ne spėk.
Man geriau spraga nei užtikrintas klaidingas atsakymas.

Kai esi tikras, atsakyk įprastai. Kai nesi tikras, aiškiai pasakyk, kad nesi tikras, ir paaiškink, ko reikėtų, kad rastume tikrą atsakymą. Neužpildyk spragų spėjimu, kuris skamba užtikrintai.

04

Promptas 3: patikra naujame pokalbyje

Stipriausias iš trijų. Atsidaryk visiškai naują pokalbį, įklijuok gautą atsakymą ir paprašyk šviežio AI jį išnarstyti. Jokios istorijos, jokios ištikimybės pirmam atsakymui, tik švarios antros akys.

Kodėl svarbu, kad pokalbis būtų naujas: tame pačiame pokalbyje AI linkęs ginti savo ankstesnį atsakymą, o ne gaudyti klaidą. Jis jau įsipareigojo, tai ir laikosi. O naujas pokalbis to atsakymo nerašė, todėl neturi ko saugoti ir daug noriau randa skyles.

Naudok viskam, kas svarbu. Atsidaryk naują pokalbį (dar geriau kitą AI: Claude atsakymą įklijuok į ChatGPT arba atvirkščiai), įklijuok atsakymą, kurį nori patikrinti, ir po juo šitą promptą.

Štai atsakymas iš kito AI. Būk skeptiškas. Surask klaidas ir patikrink, ar kiekvienas šaltinis tikrai pagrindžia teiginį, prie kurio pridėtas. Išvardink viską, kas klaidinga arba nepatikrinama.

Nelaikyk jo teisingu. Tavo darbas yra pagauti klaidas, o ne sutikti.

Gražink man:

1. Viską, kas faktiškai klaidinga, ir teisingą versiją, jei ją žinai.
2. Kiekvieną šaltinį, kuris iš tiesų nepagrindžia teiginio šalia jo.
3. Viską, ko negalima patikrinti ir kuo dar nereikėtų pasitikėti.

[ĮKLIJUOK PIRMĄJĮ ATSAKYMĄ]

05

Sąžiningai: gerėja, bet kol kas tikrini tu

Su kiekviena nauja modelių versija haliucinacijų mažėja, ir naujesni

@APIX.STUDIJA · APIXSTUDIJA.LT/GIDAI/3-PROMPTAI-KAD-AI-NEPRIGALVOTU/ · NEMOKAMAS GIDAS,
modeliai dažniau pasako „nežinau“. Tai tikra problema, prie kurios kūrėjai

rimtai dirba, ir ji traukiasi. Bet ji dar neišspręsta, todėl kol kas atsakomybė patikrinti svarbius faktus lieka tau.

Paprasta taisyklė: jei rizika maža (idėjų lietus, juodraštis), tiesiog daryk. Jei rizika didelė (skelbsi, siųsi, priimsi sprendimą), pirma paleisk patikrą naujame pokalbyje. Ji stipriausia ne be reikalo.

06

Ką padaryti šiandien

Surask paskutinį AI atsakymą su skaičiumi ar data, kuriuo jau pasinaudojai, ir paleisk jam trečią promptą naujame pokalbyje. Pamatysi, kiek jame išlieka.