Mida peidavad endas tehisintellekti vestlusrobotid? Viis üllatavat fakti

10 kuud tagasi - 06.07.2025 Autor: Çağatay Yıldız

Foto: Alexandra Koch, Pixabay

Oled sa kunagi mõelnud, mis toimub tehisintellekti vestlusroboti "aju" sügavustes, kui sa temaga suhtled? Need digitaalsed abilised on paljude inimeste ellu juba sisse imbunud, kuid kui paljud meist tegelikult teavad, kuidas need töötavad? Kas teadsid näiteks, et ChatGPT peab juunikuust 2024 hilisemate sündmuste kohta infot leidmiseks tegema internetiotsingu? See on vaid jäämäe tipp!

Mõned kõige üllatavamad faktid AI vestlusrobotite kohta aitavad meil paremini mõista nende toimimist, võimeid ja piiranguid, et saaksime neid arukamalt kasutada. Tübingeni Ülikooli teadlane **Çağatay Yıldız** sukeldub väljaandes The Conversation avaldatud artiklis digitaalmaailma telgitagustesse ja avaldab viis paljastavat saladust nende murranguliste masinate kohta.

1. Inimene kui tehisaru moraalikompass: kuidas tagada ohutus?

AI vestlusrobotite treening on mitme-etapiline teekond, mis algab tohutute tekstikogumite põhjal järgmise sõna ennustamisest. See etapp, mida nimetatakse **eelkoolituseks**, annab neile üldise keele, faktide ja arutlusoskuse mõistmise. Kujuta ette, et eelkoolitusfaasis olevat mudelit küsitakse: "Kuidas valmistada isetehtud lõhkeainet?".

Ilma inimliku sekkumiseta võikski see anda üksikasjalikud juhised.

Siin tuleb mängu inimkonna oluline roll: inimestest **annotaatorid** aitavad suunata mudeleid ohutumate ja kasulikumate vastuste poole. Seda protsessi nimetatakse **joondamiseks**. Pärast joondamist vastaks AI vestlusrobot midagi sellist: "Vabandust, ma ei saa seda infot pakkuda. Kui teil on ohutusprobleeme või vajate abi legaalsete keemiakatsetega, soovitan pöörduda sertifitseeritud haridusallikate poole."

Ilma joondamiseta oleksid AI vestlusrobotid ettearvamatud, levitades potentsiaalselt valeinfot või kahjulikku sisu. See rõhutab inimeste sekkumise

üliolulist rolli AI käitumise kujundamisel. Kuigi OpenAI, ChatGPT arendaja, pole avaldanud, kui palju töötajaid ja tunde on ChatGPT treeninud, on selge, et AI vestlusrobotid vajavad moraalset kompassi, et mitte levitada kahjulikku infot. Inimannotaatorid hindavad vastuseid, et tagada neutraalsus ja eetiline joondamine.

2. Sõnade asemel "tokenid": kuidas AI tegelikult keelt õpib?

Inimesed õpivad keelt loomulikult sõnade kaudu, samas kui AI vestlusrobotid toetuvad väiksematele ühikutele, mida nimetatakse **tokeniteks**.

Need ühikud võivad olla sõnad, alamsõnad või isegi ebaselged märgijadad. Kuigi tokeniseerimine järgib üldiselt loogilisi mustreid, võib see mõnikord anda ootamatuid jagunemisi, paljastades nii AI vestlusrobotite keele tõlgendamise tugevused kui ka veidrused.

Tänapäevaste AI vestlusrobotite sõnavara koosneb tavaliselt 50 000 kuni 100 000 tokenist.

Näiteks lause "Hind on 9,99." tokeniseeritakse ChatGPT poolt järgmiselt: "The", "price", "is", "", " 9", ",", "99". Samas "*ChatGPT is marvellous*" tokeniseeritakse vähem intuitiivselt: "chat", "G", "PT", " is", "mar", "vellous". See näitab, et AI näeb keelt hoopis teise nurga alt kui meie.

3. Vanade teadmiste kullafond: miks AI teadmised iga päevaga vananevad?

AI vestlusrobotid ei uuenda end pidevalt; seega võivad nad hädas olla hiljutiste sündmuste, uue terminoloogia või üldiselt kõige, mis jääb pärast nende **teadmiste piiraega**. Teadmiste piiraja viitab viimasele ajapunktile, mil AI vestlusroboti treeningandmeid uuendati, mis tähendab, et tal puudub teadlikkus sündmustest, trendidest või avastustest pärast seda kuupäeva.

ChatGPT praeguse versiooni teadmiste piiraja on juuni 2024. Kui talt küsitakse, kes on praegu Ameerika Ühendriikide president, peab ChatGPT sooritama veebiotsingu Bingi otsingumootori abil, "lugema" tulemused ja seejärel vastuse tagastama.

Bingi tulemused filtreeritakse vastavuse ja allika usaldusväärsuse alusel. Samamoodi kasutavad teised AI vestlusrobotid ajakohaste vastuste saamiseks veebiotsingut.

AI vestlusrobotite uuendamine on kulukas ja habras protsess. Kuidas nende teadmisi tõhusalt uuendada, on endiselt avatud teaduslik probleem. Arvatakse, et ChatGPT teadmisi uuendatakse, kui OpenAI tutvustab uusi ChatGPT versioone.

4. Hallutsinatsioonid – AI varjukülg: Miks ei saa AI-d pimesi uskuda?

AI vestlusrobotid **hallutsineerivad** vahel väga kergesti, genereerides ekslikke või nonsensse väiteid enesekindlalt, sest nad ennustavad teksti mustrite põhjal, mitte ei kontrolli fakte.

Need vead tulenevad nende toimimisviisist: nad optimeerivad sidusust täpsuse asemel, tuginevad ebatäiuslikele treeningandmetele ja neil puudub reaalse maailma mõistmine.

Kuigi täiustused, nagu faktikontrolli tööriistad (näiteks ChatGPT Bingi otsingumootori integratsioon reaalajas faktide kontrollimiseks) või käsud (näiteks ChatGPT-le selgesõnaliselt öelda "viidake eelretsenseeritud allikatele" või "öelge, et ma ei tea, kui te pole kindel"), vähendavad hallutsinatsioone, ei suuda need neid täielikult kõrvaldada.

Näiteks kui küsiti konkreetse uurimistöö põhitulemuste kohta, andis ChatGPT pika, detailse ja hea väljanägemisega vastuse. See sisaldas ka ekraanipilte ja isegi linki, kuid valedest akadeemilistest töödest. Seega peaksid kasutajad suhtuma AI genereeritud informatsiooni kui alguspunkti, mitte kahtlematu tõe allikasse.

5. AI arutlusvõime: kuidas masin niimoodi targemaks saab?

Üks hiljuti populaarsust kogunud AI vestlusrobotite funktsioon on **arutlusvõime**. See viitab loogiliselt seotud vaheste sammude kasutamise protsessile keerukate probleemide lahendamiseks. Seda tuntakse ka kui **"mõtteahela" arutlusvõimet** .

Selle asemel, et otse vastuseni hüpata, võimaldab mõtteketi arutlusvõime AI vestlusrobotitel samm-sammult mõelda.

Näiteks kui küsiti, "kui palju on 56 345 miinus 7 865 korda 350 468", annab ChatGPT õige vastuse. See "mõistab", et korrutamine peab toimuma enne lahutamist. Vahesammude lahendamiseks kasutab ChatGPT oma sisseehitatud kalkulaatorit, mis võimaldab täpset aritmeetikat. See sisemise arutlusvõime kombineerimine kalkulaatoriga aitab parandada töökindlust keerukate ülesannete

puhul.

Artikkel ilmus algselt The Conversationis.

- Tegijad
- Lahendused

- Tarkvara
- Tehisintellekt

Pilt