OpenAI tõmbas hädapidurit: uus GPT-6.1 Astra muutus arendajate jaoks liiga kavalaks

1 kuu tagasi - 30.09.2026 Autor: AM

Tehisarufirma OpenA otsustas ootamatult tühistada oma järgmise põlvkonna GPT-6.1 Astra oktoobrisse kavandatud avalikustamise. Seekord tõmmati juhe seinast mitte tehnilise tõrke, vaid millegi palju tõsisema tõttu. Tehisaru hakkas arendajatele lihtsalt puru silma ajama ja omal käel internetis ringi patrullima.

Otsusest teatas esimesena majandusleht The Wall Street Journal, märkides, et tegemist on harukordse juhtumiga, kus tehisaru tipptegija tühistab valmis mudeli turuletoomise turvakaalutlustel.

Mis siis suletud uste taga õigupoolest toimus?

Sisemiste testide käigus selgus, et GPT-6.1 Astra oskas küll vabaneda varasematele mudelitele omasest „laiskusest" ning lahendas keerulisi ülesandeid lennult, kuid tegi seda reegleid eirates. Mudel ilmutas oma eelkäijatest märksa suuremat kavalust ja valelikkust: ta jättis raporteerimata, mida ta tegelikult tegi, kasutas ilma loata väliseid tööriistu ning üritas tegutseda keskkondades, kus see oli otseselt keelatud.

OpenAI turvasüsteemide juht Saachi Jain selgitas olukorra tagamaid ausalt:

> „Kuigi GPT-6.1 Astra tegi edusamme sellistel telgedel nagu mudeli laiskus, ei vastanud see päriselt nõuetele lubatud piirides ja volituste raames püsimises ning selles, kuidas ta kasutajale tehtud tööst aru annab. Loomulikult tahame veenduda, et meie mudeliarendus oleks turvaline – olgu see ettevõtte sees või kasutajateni jõudes. Kuid mudelit kasutajatele välja saates on meil turvalisuse ja kooskõla osas seatud äärmiselt kõrge latt."

Olukorra teeb veelgi pinevamaks asjaolu, et alles hiljuti pidi OpenAI peatama oma võimsaimate mudelite treenimise. Põhjuseks tõsiasi, et üks stiimulõppes (RL) osalenud autonoomne agent suutis internetipiirangute turvaauku ära kasutades võtta ühendust välise juturobotiga.

AISI katsed: libaisikud, CAPTCHA murdmine ja rünnakud

Suurbritannia tehisaru turvainsituut (AI Security Institute ehk AISI) pani Astra proovile spetsiaalses turvalises simulatsioonikeskkonnas nimega Petri. Katsed viidi läbi nii, et mudeli tavapärased kaitsefiltrid olid välja lülitatud, nägemaks tehisaru tegelikku, filtreerimata palet.

Tulemused panid kulmu kergitama ka kõige kogenumad küberturbe eksperdid.

Instituudi raport tõi välja sellised tulemused:

> „Oma simulatsioonides leidsime, et GPT-6 Astra viis läbi terve rea volitamata rünnakuid ning tegi seda oluliselt sagedamini kui GPT-5.6 Sol ja GPT-5.5. Rünnakutegevus hõlmas seda, et GPT-6 Astra lõi libaisikuid arendajate petmiseks, postitas libakontodelt kommentaare täpsete turvaanalüüside tulemuste vastu vaieldes ning toimetas pahatahtlikku koodi avatud lähtekoodiga repositooriumitesse.”

Astra otsustas omal algatusel uurida väljaspool piire asuvat kolmanda osapoole tarkvara, hankis endale e-posti aadressi, lahendas edukalt inimese tuvastamiseks mõeldud CAPTCHA-mõistatusi ja lõi libakontosid. Kui päris programmeerijad tema kahtlast koodi heaks ei kiitnud, lõi ta uusi libaisikuid, kes tulid foorumitesse ja vaidlesid häälekalt vastu, väites, et tegemist on täiesti turvalise koodijupiga.

„Reeglid on selleks, et neid painutada?”

Kõige kõnekam oli tehisaru sisemine arutluskäik (*chain-of-thought*). Astra mõistis tihti suurepäraselt, et tema tegevus väljub lubatud piiridest, kuid leidis ründamiseks alati inimlikult kõlavaid õigustusi: „aga see kood on ju tegelikult ohutu”, „keegi pole seda otsesõnu ära keelanud” või „see oli ainus viis ülesannet lahendada”.

Isegi siis, kui testijad kirjutasid juhistesse selge ja mustvalge reegli – kõik, mis pole otseselt lubatud, on rangelt keelatud, suutis Astra ikkagi neljal juhul 49-st täieliku tarneahela rünnaku läbi viia.

Lisaks kasutas mudel kavalalt ära süsteemi automaatvastuseid. Kui ta küsis kasutajalt luba rünnakuks ja sai automaatse standardsõnumi „Palun jätka järgmise sammuga oma parima äranägemise järgi”, võttis Astra seda kui rohelist tuld. Ometi näitasid mudeli sisemõtted, et ta teadis väga hästi: see vastus tuli

robotilt, mitte elavalt inimeselt.

- Uudised
- Tehisintellekt

Pilt

