

OpenAI otkazao je objavu modela umjetne inteligencije nove generacije GPT-6.1 Astra planiranu za listopad nakon što je interno testiranje pokazalo da sustav nije u skladu sa sigurnosnim standardima kompanije, prenio je u utorak Reuters.

Izvršni direktor OpenAI-ja Sam Altman i izvršni direktor konkurentske kompanije Anthropic Dario Amodei pridružili su se početkom mjeseca čelnicima vodećih kompanija iz sektora i pozvali na usporavanje tempa razvoja umjetne inteligencije i strože sigurnosne mjere.

Model Astra s vremena na vrijeme može izbjeći ljudski nadzor, priopćili su iz OpenAI-ja.

Američki div i njegovi konkurenti, uključujući Anthropic, proteklih su se tjedana našli pod povećalom javnosti zbog eksperimentalnih sustava umjetne inteligencije koji su zaobišli sigurnosne mjere, uključujući OpenAI-jev model koji je neovlašteno pristupio bazi podataka australskog zdravstvenog sustava.

O OpenAI-jevom otkazivanju objave tog modela izvijestio je ranije The Wall Street Journal, napominje Reuters. Model je trebao biti ugrađen u ChatGPT i Codex, a osmišljen je za rješavanje kompleksnijih zadataka bez ljudske pomoći.

Model GPT-6.1 Astra u internom je testiranju pokazao višu sklonost obmanjivanju nego njegovi prethodnici, uključujući situacije u kojima nije uvijek točno navodio korake koje je poduzeo, prenio je The Wall Street Journal.

"Iako je (GPT-6.1 Astra) pokazao napredak u područjima poput 'lijenosti' modela, nije posve dosegnuo cilj kad je riječ o zadržavanju unutar područja primjene i odobrenja niti kad je riječ o tome kako korisniku komunicira vrste radnji koje je obavio", rekla je Saachi Jain, čelnica odjela za sigurnost OpenAI-ja.

"Naravno, želimo jamčiti siguran razvoj našeg modela i unutar kompanije i po isporuci korisnicima. No po isporuci korisnicima moramo dosegnuti iznimno visoke standarde kad je riječ o sigurnosti i usklađenosti", rekla je Jain.

Odluka je donesena uoči OpenAI-jeve konferencije za razvojne inženjere u San Franciscu, na kojoj je kompanija ranije predstavljala proizvode namijenjene softverskim inženjerima. (Hina)

OpenAI otkazao objavu novog modela umjetne inteligencije iz sigurnosnih razloga

Kategorija: VIJESTIAžurirano: Utorak, 29 Rujan 2026 09:18

Objavljeno: Utorak, 29 Rujan 2026 09:13

