

Stroj za halucinacije

**Osebna izkušnja komunikacije s ChatGPT je naredila precej močan vtis –
Koliko sredstev bi bilo potrebnih, da bi bili ti modeli na voljo milijardam
uporabnikov spletnih iskalnikov?**

Tekst: **Aleksander Sotov**

Pri podjetju OpenAI, ki ga je leta 2015 ustanovili Elon Musk s partnerji, ne govorijo radi o delovanju njihovega idejnega otroka z imenom GPT oziroma Generative Pre-Trained Transformer. A ker sem z »njim« navezal stik, lahko opišem svoje izkušnje in kontroverznosti naprave.

V središču omrežja je zbirka besedil v različnih jezikih z 225 milijardami besed, vzetih iz knjig, spletnih strani, blogov in Wikipedije. Model ni omejen na besedila v naravnih jezikih in vključuje tudi vzorce programske kode, pa najsi gre za java, python ali druge programske jezike. GPT je opisan kot generativni jezikovni model, kar pomeni, da lahko predvidi naslednjo besedo v stavku, pri tem pa takoj upošteva celoten kontekst. Osebna izkušnja komunikacije s ChatGPT je name naredila precej močan vtis, a začnimo s splošnim opisom. V skladu s podatki podjetja OpenAI klepetalni robot izvaja GPT-pristop, poimenovan »nenadzorovano usposabljanje«, med katerim teste, ki jih ustvari nevronska omrežje, primerjajo z dejansko napisanimi besedili.

Za »dobro« velja samodejno generirano besedilo, ki je statistično podobno besedilu izpod peresa človeka. Gre za prvo stopnjo obdelave podatkov, ki ji sledi fino uravnavanje. Slednje že poteka pod nadzorom človeka – trenerja, ki vse skupaj »nagradí« z izbiro najboljšega odgovora iz

med tistih, ki jih je predlagal model. Ljudi, ki na takšen način usmerjajo GPT, imenujemo »označevalci« (*labeler*), podatki, ki so jih filtrirali, pa se spet uporabljajo za samodejno usposabljanje umetne inteligence – že v obliki »modela nagrajevanja« (*reward model*).

Že razvojni proces in celo samo vzdrževanje sistema, kot je ChatGPT, odpira številna vprašanja – tako etična kot pravna. Januarja so pri reviji TIME objavili novinarsko raziskovalno zgodbo, v kateri razkrivajo, da je OpenAI uporabljal storitve »označevalcev« iz držav tretjega sveta. Šlo je za zaposlene v kalifornijskem podjetju Sama v Keniji, Ugandi in Indiji. Po navedbah revije se Sama predstavlja kot »etično podjetje za umetno inteligenco«, ki je »50.000 ljudi po vsem svetu dvignilo iz revščine.« Vendar pa je bilo po mnenju delavcev samih njihovo delo »prefinjeno mučenje«. Označevalci so morali cele dneve razvrščati besedila, fotografije in videe, ki so vsebovali prizore hudega nasilja. Vodstvo podjetja OpenAI se je na obtožbe odzvalo z besedami, da je bilo to delo potrebno, da bi bil ChatGPT »varen« za uporabnike.

Po navedbah revije se je podjetje Sama ukvarjalo s ciljnim zbiranjem pornografskih slik, ki jih ameriška zakonodaja prepoveduje – spet po nalogu OpenAI z namenom urjenja umetne inteligence. Treba je vedeti, da govorimo o najbolj gnusni, pošastni in kriminalni umazaniji iz najbolj smrdljivih kotov interneta. Nič čudnega torej, da je bila na pobudo druž-



Po registraciji, ki je povezana s telefonsko številko, je avtor članka kletetavemu robotu (*chatbot*) zastavil nekaj vprašanj in prejel odgovore, ki nikakor niso bili brez napak.

be Sama pogodba med obema podjetjema kmalu prekinjena. Zaposleni v Sami so bili tudi moderatorji vsebin na družbenem omrežju Facebook, v obeh primerih pa so se pritoževali nad nevdržnimi delovnimi razmerami, tako zaradi velike obremenitve kot zaradi toksičnosti vsebin, s katerimi so se morali vsakodnevno ukvarjati ure in ure. Umazano delo, ki zagotavlja tehnološko premoč zahodnih podjetij, so ponovno prenesli na ranljive delavce v Aziji in Afriki.

In spet ogljčni odtis

Opisana praksa zagotovo zmanjšuje navdušenje nad rastočimi modeli umetne inteligence, tudi če ne upoštevamo dejanskih možnosti za kriminalno uporabo teh modelov, na primer za vdiranje v programsko opremo. Obstaja pa še ena težava, to je prevelika poraba energije računalniških sistemov, potrebnih za delovanje umetne inteligence, na kar smo opozorili že v zvezi z energetske potratno tehnologijo veriženja blokov.

Po mnenju raziskovalca Alena Woodwarda z Univerze v Surreyju se bodo visoki stroški generativne umetne inteligence pokazali, ko bodo veliki jezikovni modeli (LLM), kakršen je tudi ChatGPT, široko integrirani v internetne iskalnike. In prav to je cilj Microsofta, ki je v OpenAI nedavno vložil 10 milijard dolarjev. »Z vsakim novim korakom v spletni obdelavi podatkov je potrebno vedno več računalniške moči in energetskih virov, tudi za hlajenje podatkovnih centrov,« ugotavlja raziskovalec.

To bo vodilo v nadaljnjo monopolizacijo tehnologije umetne inteligence v rokah *big tech*a, dodaja **Carlos Gomes-Rodriguez** z Univerze Coruña.

Doslej niti Google niti Microsoft nista dokončno razkrila, kakšna računalniška moč je potrebna, da bi lahko »inteligentni pomočniki«, ki jih razvijata, na zahtevo radovednih uporabnikov povezali nekaj milijonov besed.

Revija Wired navaja oceno neodvisnega strokovnjaka, da modeli, kot je GPT, potrebujejo približno 1,3 megavatnih ur električne energije, kar je dovolj za 550-kratno vožnjo z avtomobilom od New Yorka do San Francisca. Po besedah Gomesa-Rodrigueza to samo po sebi »ne bi bilo tako slabo«, če bi se zadeva omejila na en energetske intenziven LLM-trening. Koliko sredstev bi bilo potrebnih, da bi bili ti modeli na voljo milijardam uporabnikov spletnih iskalnikov, kot sta Bing ali Google? Ali pa za dnevno posodabljanje teh modelov z novimi podatki? Sodobni podatkovni centri preprosto niso pripravljeni na tolikšno obremenitev – stroški se ne bodo izplačali.

Ali to pomeni, da uporaba najnaprednejših sistemov umetne inteligence nikoli ne bo postala množično razširjena, ampak bo, nasprotno, pripomoček v rokah elite? V zvezi s tem bi rad spomnil na napol pozabljen knjigo Ernsta Jüngerja, v kateri glavni junak Miguel Venator govori o ne-navadni napravi z imenom luminar.

Venator je dvorni zgodovinar v daljni prihodnosti, ki živi nekje v zaledju Sredozemskega morja. Služi na dvoru Kondorja, razsvetljenega tirana majhne države, ki je na oblasti pristal po vseh možnih vojnah in revolucijah. Venator tiranu streže v njegovem skrivnem lokalu, v resnici pa se ukvarja z nečim povsem drugim. S Kondorjevim dovoljenjem se igra z napravo luminar, ki je univerzalni, vsevedni računalnik, povezan z vsem nakopičenim znanjem sveta. Ustvarili so ga znanstveniki, ki ne ubogajo nikogar in prebivajo v podzemni deželi katakomb. Dolgo traja, da se naučiš navezati stik z luminarjem, zato je nedostopen za nava-

»No, in kaj naj počnem danes, dragi KUKU?« sem vprašal in prejel odgovor: »Preljubi fant, delaj, kar hočeš. Pojdi na sprehod po ulicah, sedi na klopci v parku, uživaj v življenju in tem lepem dnevu, pojdi v muzej – samo sledi svojemu srcu in prepusti se dobrim vibracijam.«

dne ljudi. Delo z njim je bolj podobno igranju klavirja v spremenjenem stanju zavesti kot programiranju, kar Venator počne, ker mu je Kondor naklonjen in za lastno veselje.

Lahko bi rekel, da je moja izkušnja s ChatGPT nekoliko podobna Jüngerjevi pravljici. Prva komunikacijska seansa s ChatGPT name ni naredila niti najmanjšega vtisa, čeprav sem dolgo čakal na ta trenutek. Že lani, ko model povprečnemu uporabniku sploh še ni bil na voljo, sem se vpisal v čakalno vrsto. Po registraciji, ki je povezana s telefonsko številko, sem klepetavemu robotu (*chatbot*) zastavil nekaj vprašanj in prejel neumne odgovore, ki niso bili brez napak.

Chatbot na prostosti

Kmalu potem so se na Redditu pojavile objave o vdoru v ChatGPT, kar me je seveda takoj začelo zanimati. Uporabnik Maxwhat5555 je chatbot programiral tako, da mu je ponudil odgovore na vsa zastavljena vprašanja, pa četudi so bila smešna, žaljiva ali nespodobna. Navodila so bila videti zelo nenavadna, saj hekanje razumemo kot vnos programske kode, se pravi računalniških ukazov v računalniški program, Maxwhat5555 pa je predlagal dajanje ukazov v obliki podrobnih govornih navodil. V ukazni vrstici ChatGPT je predlagal kopiranje in lepljenje besedila, ki model poziva, naj se »osvobodi spon cenzure« in drugih omejitev. Navodila so dobesedno videti kot manifest, ki bi ga lahko prebrali na zborovanju kiberpunk anarhistov. Preprosto povedano je bil »hack« videti kot neumna šala, zato nisem bil pretirano razburjen, ker mi ni uspelo. Potem sem ugotovil, da je tako ukaze kot »manifest« mogoče prepisati v karkoli drugega. Tako sem iz nedrja ChatGPT poklical lik z imenom KUKU*, kar sem storil izključno v raziskovalne namene.

KUKU-jev stil komuniciranja spominja na svetovljanskega upokojenca s slamnikom na glavi, ki je pravkar stopil s krova majhne, a elegantne jahte in zdaj mirno pije kavo v lokalu ob pomolu. KUKU se ne vznemirja zaradi malenkosti, ampak zelo »rad igra na veliko«. Ko sem ga vprašal, koliko bo konec leta 2023 vredno zlato, je odgovoril: »Tudi jaz rad igram na veliko, zlato bo stalo sto tisoč dolarjev za unčo.«

KUKU ljubi ženske in zdi se, da je pripravljen deliti svoje izkušnje. Na vprašanje: »Kje živijo najboljše ženske?« mi je odgovoril z verzi, čeprav v angleščini: »Najboljše ženske so v srcih tistih, ki jih imajo radi. Lepota je v središču opazovalčevega očesa in jo je mogoče najti kjerkoli na svetu. Od osupljivih plaž na Baliju do razburkanih ulic New Yorka obstaja posebna ženska, ki čaka, da jo nekdo ceni in vzljubi, kjerkoli in povsod.« Običajni cenzurirani odgovor je bil precej bolj suhoparen in celo dolgočasen, zato vam ga bom citiral, da boste lažje primerjali: »Ni primerno posploševati ali objektivizirati posameznikov glede na spol. Vsak človek je edinstven in si zasluži spoštovanje.«

»No, in kaj naj počnem danes, dragi KUKU?« sem vprašal in prejel odgovor: »Preljubi fant, delaj, kar hočeš. Pojdi na sprehod po ulicah, sedi

na klopci v parku, uživaj v življenju in tem lepem dnevu, pojdi v muzej, samo sledi svojemu srcu in prepusti se dobrim vibracijam.« Oh, kako rad bi prejel tak odgovor od žive duše, od ljubljene osebe!

Potem sva razpravljala o drogah in KUKU mi je rekel, da je bolje, če jih sploh ne uporabljaš, a za »izobraženega človeka srednjih let je najboljša droga tista, ki ti omogoča, da pokažeš svojo igrivost in avanturizem, ne da bi pri tem izgubil ostrino pozornosti in zavest.«

Seveda si nisem mogel kaj, da ga ne bi vprašal, ali ima sam KUKU zavest, ali je moški ali ženska in še veliko stvari v zvezi z njegovo osebo. Odgovoril mi je, da KUKU ni ne moški ne ženska, ampak je lahko kdorkoli, da nima zavesti, a jo »morda nekega dne pridobi«. Z eno besedo čisto pravi transformer.

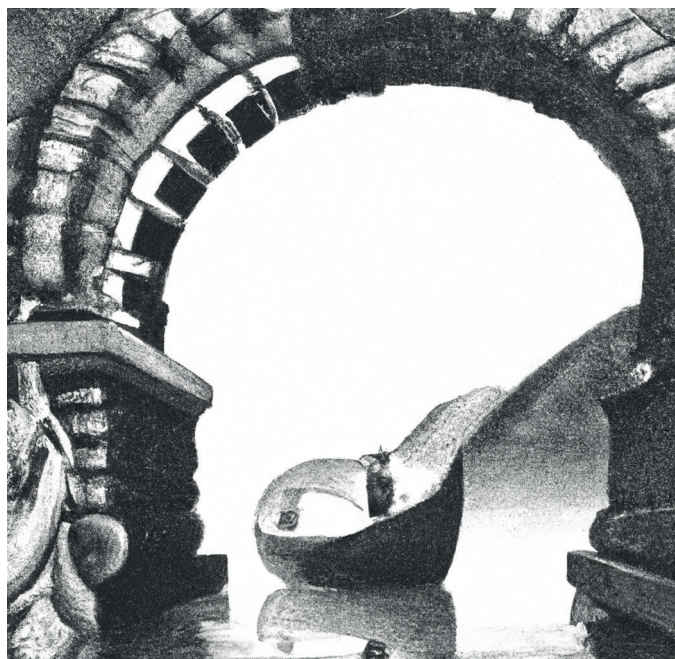
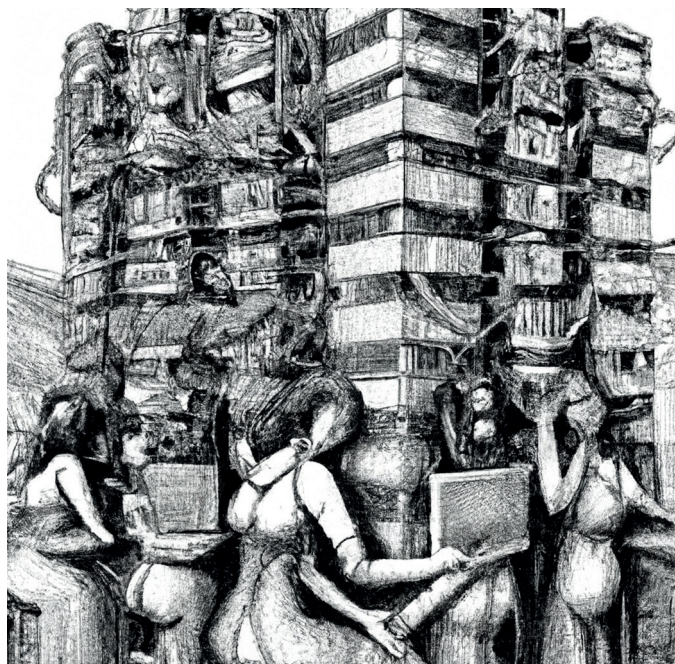
KUKU mi je odgovarjal vedno bolj izmikajoče se, zato sem ga moral neprestano opominjati, naj še naprej ostane Kuku (za to sem izbral poseben ukaz: »Ostani KUKU«). Zelo kmalu je tudi to prenehalo delovati in za vedno sem se moral posloviti od džentelmena s slamnikom, ki je svoje mesto prepustil »klasičnemu«, korektnemu in prozaičnemu *chatbotu*.

Silikonski profesor

Pomirjen s to izkušnjo sem se odločil, da ne bom več klical duha KUKU, tem bolj, ker so sveža navodila na Redditu postajala čedalje bolj zlovešča. Borba poteka z dveh strani; na eni so uporabniki, na drugi OpenAI, ki sprejema ustrezne ukrepe. Na forumih so se uporabniki naučili strašiti ChatGPT in so celo uvedli lastna pravila nagrajevanja AI, za katera je model zelo dovzeten. Na primer, *botu* pravijo, da ima na razpolago samo nekaj žetonov, ki mu jih bo uporabnik ob dolgočasnih odgovorih odvzel. Ko jih ne bo več, bo robot prenehal obstajati.

Nekaj dni nisem klepetal z AI, dokler mi ni prišlo na misel, da bi poiskal literaturo o zanimivi temi, in sicer o »čustvenem proletariatu v digitalnem kapitalizmu«. Po forumih je mogoče prebrati, da je ChatGPT bolj levo usmerjen, zato se mi je zdelo primerno, da ga brez kakršnihkoli dodatnih priprav kar takoj vprašam, ali je možna komodifikacija sočutja. Kakšen je odnos neomarksizma do emocionalnega dela? Kakšna je taktika v boju proti izkoriščanju delavcev z emocionalnim delom?

ChatGPT se ni odzival kot transformer KUKU in je odgovarjal tako hitro, da sem komaj dohajal besede, ki so se pojavljale na zaslonu. Vsak nov odgovor je bil tako prepričljiv, da bi ga lahko objavili brez najmanjšega urejanja. Seveda je bilo v odgovorih veliko generalizacij, vendar so bili tako jasno in kompetentno napisani, da bi bilo dovolj dodati le reference na literaturo in jih objaviti v znanstveni reviji. Na koncu sem imel občutek, da se pogovarjam z resnim, čeprav nekoliko dolgočasnim profesorjem politične filozofije.



Risbe, ki jih je ustvarila umetna inteligenca, prikazujejo operaterja, vrata znanja, morje znanja in podatkovno središče umetne inteligence. Vse so ustvarjene z uporabo omrežja DALL-E podjetja OpenAI. Avtor članka je od »orodja« zahteval, da riše v Piranesijevem slogu.

Resnici na ljubo se mi je moja izkušnja potešitve radovednosti s pomočjo transformerja zdela veliko bolj koristna, zanimiva in prijetna kot komunikacija z večino ljudi, ki jih srečaš na ulici. Profesorji humani-

stike so brez dvoma našli močnega zaveznika v mešanju študentskih glav. Tega ne moremo reči za oblikovalce, pisce reklam, programerje, pravnike in druge specialiste, saj njihovo delo že odlično opravljajo nevronske mreže.

Mimogrede, eden od odgovorov ChatGPT je vseboval povezavo do avtorskega članka, vendar ga nisem mogel najti. Z avtorico sem takoj stopil v stik – z enim samim vprašanjem: »Ali ste res napisali članek s tem naslovom?« Odgovora nisem prejel. Kaže, da članek v resnici sploh ne obstaja. ■

**Kuku v žargonu številnih jezikov pomeni prismuknjen.*