Halucinacije v generativnih inteligentnih sistemih (GPT) – napaka zasnove ali znak porajajoče se inteligence?

Jurij Bon, psihiater in nevroznanstvenik

Običajno je sposobnost kompleksne kognicije, oziroma višjih živčnih procesov, ki so podlaga inteligentnosti, razumljena kot lastnost višje razvitih bioloških organizmov. Umetni inteligentni sistemi, v zadnjem času so najbolj razvpiti generativni, ki jih poznamo tudi pod kratico GPT, izkazujejo nepričakovano kompleksnost pri ustvarjanju besedil ali podob. Razviti so bili kot sistemi za prepoznavanje vzorcev na podlagi statističnih razmerij med podatki, na katerih se učijo. Ker imajo dostop do množice jezikovnih in drugih podatkov na svetovnem spletu, na katerih se naučijo omenjenih statističnih razmerij v vzorcih besedil, in jih podpira velika strojna računska moč, so sposobni generirati besedila ali podobe zavidljive kompleksnosti, čeprav se v osnovi ne zavedajo oziroma razumejo pomena svojih izdelkov. Tako kompleksno vedenje nas hitro lahko zavede, da jim pripišemo kognitivne sposobnosti, oziroma celo primitivno samozavedanje, za kar pa opozarjajo, da je lahko posledica neustrezne antropomorfizacije, iskanja človeških lastnosti v umetnih računskih sistemih [1]. Šlo naj bi enostavno za statistično preračunavanje, katere besede najverjetneje sledijo predhodnemu vzorcu besed, kar le na površini in zaradi ogromne množice upoštevanih podatkov daje vtis inteligentnega, kognitivnega procesa. Oziroma z drugimi besedami, »generativni jezikovni modeli ne razumejo sveta in odnosov v njem, se ne zavedajo da jim nasproti stojijo inteligentna in zavestna človeška bitja, ampak se preprosto ukvarjajo z inteligentnim ugibanjem, kako se množica besed najbolj verjetno povezuje med seboj« [2].

Kdorkoli, ki je v zadnjem času pogosteje uporabljal na spletu dostopne generativne sisteme, med katerimi je najverjetneje najbolj poznan jezikovni model ChatGPT, se je lahko srečal tudi s tako imenovanimi halucinacijami, ko sistem poda navidez smiselne odgovore, ki pa so lahko delno netočni ali pa popolnoma izmišljeni. Take halucinacije običajno interpretiramo kot napako v procesu generacije vsebine, kar izhaja iz razumevanja, da naj bi generativni sistemi delovali predvsem kot pomočniki, ki imajo dostop do množice dejstev in znanja, ki bi jih morali ustrezno predstaviti in oblikovati v pomoč človeškemu uporabniku, ki od njih pričakuje nezmotljivost. Kot možnosti popravljanja teh napak in izogibanja haluciniranju sedaj svetujejo predvsem prečiščevanje izvornega materiala, na katerem se generativni model uči povezav v vzorcih besed, ali pa nastavljanje različnih parametrov, ki določajo njegovo delovanje. Lahko bi na primer omejevali njegovo generiranje vsebin v primerih, ko je bilo povezanega izvornega materiala le malo in so izračunana razmerja manj zanesljiva, ali pa omejevali njegovo stopnjo kreativnosti, s čimer razumemo obseg vključevanja naključnosti oziroma ugibanja pri generiranju vsebin [3]. Vendar je s tem v zvezi zanimivo, da veliki jezikovni generativni sistemi, ki imajo dostop do bolj kvalitetnih in obsežnih podatkov, bolj halucinirajo kot manjši in preprostejši, kar je presenetilo strokovnjake za umetno inteligenco [3].

---

[1] Nosta J. (2023). The nature of GPT »hallucinations« and the human mind.
[2] Michalowski J. (2023). What powerful new bots like ChatGPT tell us about intelligence and the human brain.
[3] Treiber M. (2023). Why does GPT hallucinate?

Naše razumevanje, kako naj bi sistemi umetne inteligence delovali, se je skozi razvoj tega znanstvenega področja precej spremenilo. Sprva so predvidevali, da je podlaga inteligentnosti predvsem sposobnost jezikovnega procesiranja in manipuliranja simbolov, kot so besede in številke. Vendar tako razumevanje ne ustreza biološki realnosti. Organizmi lahko kažejo izrazito inteligentno vedenje, ne da bi bili sposobni uporabljati procese shranjevanja in manipulacije simbolov. Podlaga inteligentnemu vedenju v bioloških organizmih je predvsem sposobnost nevroplastičnosti, oziroma sprotnega prilagajanja moči trilijonov povezav med nevroni v živčnem sistemu. Na tem spoznanju je nato temeljil tudi razvoj sodobnih sistemov umetne inteligence, ki podobno uporabljajo pri svojem učenju kompleksnih vzorcev spreminjanje moči povezav v umetnih nevronskih mrežah [4].

V človeških možganih naj bi ravno interakcije v množici nevronskih povezav vplivale na nastanek višjih, kompleksnih kognitivnih procesov, inteligence in zavesti, ki jih biološko razumemo kot emergentno lastnost sistema. Zanimivo je, da tudi pri ljudeh lahko opazimo podobne fenomene ustvarjanja netočnosti, kot so halucinacije v generativnih sistemih umetne inteligence. Spomin, ki ga lahko ubesedimo, delimo na semantični, ki predstavlja nakopičeno znanje in povezave med pojmi, ter epizodični, v katerem se shranjujejo dogodki naše individualne zgodovine in ki je tudi podlaga našemu samozavedanju in občutku, da smo individuum, ki obstaja v toku časa. Priklic podatkov iz spomina že pri zdravih ljudeh nikdar ni natančna preslikava podatkov v spominu, ampak gre za deloma generativni proces, ki vključuje kreativnost in ugibanje. V možganih, ki so bili poškodovani ali obolijo, pa so ti fenomeni še bolj izraziti. Pri okvarah spominskih področij lahko bolniki generirajo kompleksne izmišljotine, kar je sicer bolj značilno za epizodični kot semantični spomin. Te delne ali popolne spominske izmišljotine strokovno imenujemo konfabulacije in ne halucinacije, ki nasprotno pomenijo generiranje občutka zaznav zunanjih dražljajev v zavesti v odsotnosti dejanskih dražljajev v okolici, kot so na primer neresnični glasovi pri bolnikih s psihotičnimi motnjami. Še bolj nenavadni pa so fenomeni, ko zaradi poškodbe področij, ki se primarno ukvarjajo z zaznavanjem zunanjega sveta, pride do dejanskih izpadov zunanje zaznave, ko se bolniki nenadoma nehajo zavedati, da bi te zaznave morale obstajati. Na mestih v vidnem polju, kjer so pred okvaro videli del zunanje okolice, se ne pojavijo prazna ali črna mesta, ampak se vidna zaznava zlije v novo popolno celoto, ki ne vključuje več tistih delov zunanjega sveta, ki jih zaradi okvare nismo več sposobni zaznati, čeprav v realnosti seveda še vedno obstajajo. Zanimivo je, da v takih primerih konfabulacij ali okvar zaznavanja bolniki izgubijo tudi uvid, da je z njihovim spominom ali zaznavo nekaj narobe. Kaže, da je pomembna lastnost človeške inteligence in zavesti, da na osnovi dostopnih informacij generiramo najbolj verjetno sliko realnosti, ki je za posameznika tudi edina resnična. Konfabuliranje v tem primeru torej ne pomeni nujno napake, ampak gre lahko za odraz delovanja inteligentnega sistema. V skladu s tem nekateri pomembni raziskovalci umetne inteligence menijo, da tudi haluciniranje velikih generativnih sistemov umetne inteligence morda ni napaka, ampak odraz primitivne, porajajoče se inteligence drugačne vrste, ki lahko pomeni prve korake v nadaljnjem razvoju teh sistemov v smeri prave avtonomne in samozavedajoče se umetne inteligence [4].

[4] Heaven WD. (2023). Geoffrey Hinton tells us why he's now scared of the tech he helped build.