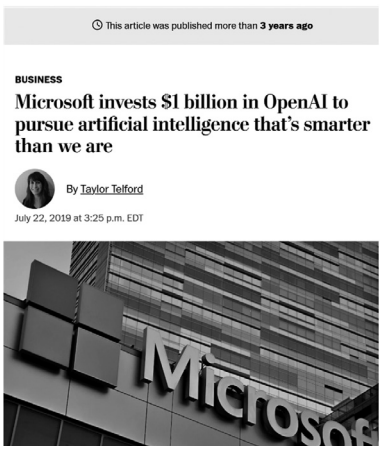
Jezični modeli[1]

Tvrtko Tadić[2]

Krajem 2022. godine lansiran je *ChatGPT*, tada široj javnosti malo poznate tvrtke *Open AI*. Proizvod je doživio neviđeni uspjeh, a njime je počelo doba umjetne inteligencije. Velike svjetske IT kompanije tom su bačenom rukavicom ušle su u utrku za najbolje razvijeni proizvod, pa se tako svaki dan najavljuju novi i bolji jezični i drugi modeli umjetne inteligencije. U ovome članku dat ćemo uvid u to što su jezični modeli, kako se koriste te koje su im trenutne mogućnosti.

Pretrage i sažimanje sadržaja

Već dulje vrijeme navikli smo tražiti sadržaj pomoću tražilica. Obično bi nakon unosa nekog pojma tražilica pronašla web-stranice koje bi korisnik pregledao i utvrdio nalazi li se na njima ono što traži. Tražilice su također bile u mogućnosti dati odgovore na jednostavna pitanja, no *ChatGPT* je svojom pojavom uzdigao stvari na višu razinu. Po prvi put korisnici su dobili mogućnost postavljanja složenijih upita i dodatnih pitanja. *Chat* je bio u mogućnosti sastaviti tekst u formi i duljini u kojoj je korisnik to želio. No, imao je dvije mane:


This article was published more than **3 years ago**

BUSINESS

Microsoft invests $1 billion in OpenAI to pursue artificial intelligence that's smarter than we are

By Taylor Telford

July 22, 2019 at 3:25 p.m. EDT

Slika 1. Naslov članka u Washington Postu iz 2019. godine

- informacije kojima je *ChatGpt* raspolagao bile su dostupne zaključno do rujna 2021.
- povremeno bi generirao tekst koji je bio netočan ili potpuno izmišljen, što je fenomen poznat kao **haluciniranje**.

U veljači 2023. *Microsoft* je uz pomoć *Open AI*-a predstavio *Bing AI* koji objedinjuje mogućnosti *ChatGPT*-a i internetske tražilice. Nekoliko mjeseci nakon toga *Google* je najavio svoju verziju nazvanu *Bard*. Počela je utrka u razvoju AI proizvoda.

[1]Članak se temelji na predavanju Veliki jezični modeli: Što se sve promijenilo razvojem ChatGPT-a? u sklopu Inženjerske sekcije Hrvatskog matematičkog društva održane 25. lipnja 2023. godine u Zagrebu (snimka je dostupna u [7]) i predavanju Science and Practice of Large Langauge Models: An Industry Perspective održanom 26. travnja 2024. na 3rd Kathy Wilkes Memorial Conference u Torinu.

[2]Tvrtko Tadić, Microsoft Corporation, Redmond

Open AI

Open AI je laboratorij za razvoj umjetne inteligencije koja je sigurna i korisna za čovječanstvo. Ima dvije sastavnice od kojih je jedna da je **neprofitna organizacija**, a druga da je **društvo s ograničenom odgovornošću.** *Microsoft* je, prema medijskim natpisima, u taj posao investirao milijardu dolara 2019. i dodatnih 10 milijardi dolara 2023. godine. *OpenAI* razvio je neke od najnaprednijih modela umjetne inteligencije kao što su GPT-3 i GPT-4, koji mogu generirati prirodni jezik na temelju teksta ili slike.

Što su (veliki) jezični modeli?

Jezični modeli proučavaju se već dulje vrijeme. Prvi jezični modeli pojavili su se u 50-ih i 60-ih godina 20. stoljeća kao dio istraživanja u području strojnog prevođenja i prepoznavanja glasa, no tek su pojavom *ChatGPT*-a došli u središte pažnje javnosti. Jezični modeli su **matematičke reprezentacije prirodnog jezika** koje omogućuju strojevima da razumiju i generiraju tekst, a temelje se na **statističkim** ili **neuronskim metodama** koje uče iz **ogromnih količina** tekstualnih podataka.

Matematički okvir

Sada ćemo opisati kako izgleda matematički okvir za jezične modele. Većina iznesenog temelji se na bilješkama [12]. **Alfabet** je konačni neprazni skup

$$\Sigma = \{\sigma_1, \sigma_2, ..., \sigma_k\}$$

čije elemente zovemo **simboli**. Tekst je konačni niz simbola. **Jezični model** vjerojatnosna je razdioba $\mathbb{P}$ na skupu svih konačnih tekstova

$$\Sigma^* = \{\sigma_1, ..., \sigma_k, \sigma_1\sigma_1, \sigma_1\sigma_2, ..., \sigma_k\sigma_k, \sigma_1\sigma_1\sigma_1, ...\}.$$

$\mathbb{P}$ je model koji sadrži sve podatke o jeziku. U idućem (pojednostavljenom) primjeru vidjet ćemo kako to funkcionira u praksi.

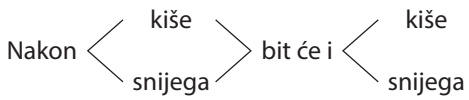
Primjer

Neka je alfabet zadan kao

$$\Sigma = \{\text{"Nakon"}, \text{"kiše"}, \text{"snijega"}, \text{"bit će i"}\}.$$

$\mathbb{P}$ možemo koncentrirati tako da vjerojatnost idućih događaja bude 1:

- Tekst ima 4 simbola
- Tekst počinje simbolom „Nakon”.
- Drugi i četvrti simbol su iz skupa $\{\text{"kiše"}, \text{"snijega"}\}$.
- Treći simbol u tekstu je „bit će i”.



Slika 2. Jednostavni jezični model

Kao što je prikazano na Slici 2.,u ovom se modelu zapravo drugi i četvrti simbol pojavljuju po nekoj vjerojatnosnoj distribuciji, dok su prvi i treći simbol zadani.

Jednostavna implementacija

Nakon što korisnik unese svoj tekst, aplikacija će:

- procijeniti vjerojatnosnu razdiobu nastavka teksta *uvjetno* na uneseni tekst, koju označavamo s

$$\mathbb{P}\left(\cdot \mid \text{korisnikov unos}\right)$$

- simulirati uzorak iz ove razdiobe i ispisati tekst korisniku

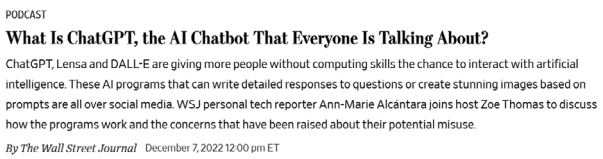
Slika 3. Shema rada jezičnih modela

Kako rade?

Nakon upisa teksta aplikacija šalje upit na *oblak* na kojemu model proizvodi odgovor te ga šalje nazad aplikaciji. U slučaju svih modela koje radi *Open AI*, oni se svi izvršavaju na *Microsoftovom* računalnom oblaku – *Azure*.

Slika 4. Shema funkcioniranja ChatGPT-a

Prema procjenama i pisanju medija, upit na *ChatGPT*-u je 4 do 7 puta skuplji od klasične pretrage (vidi Sliku 6.).

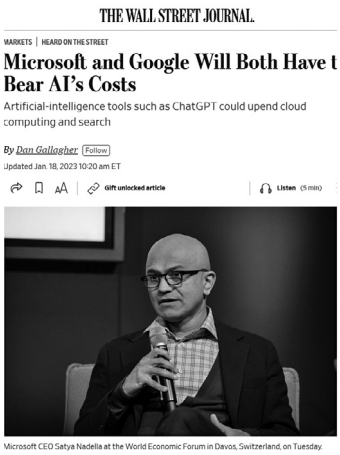

PODCAST

What Is ChatGPT, the AI Chatbot That Everyone Is Talking About?

ChatGPT, Lensa and DALL-E are giving more people without computing skills the chance to interact with artificial intelligence. These AI programs that can write detailed responses to questions or create stunning images based on prompts are all over social media. WSJ personal tech reporter Ann-Marie Alcántara joins host Zoe Thomas to discuss how the programs work and the concerns that have been raised about their potential misuse.

By The Wall Street Journal December 7, 2022 12:00 pm ET

Slika 5. Naslov podcasta kojeg je objavio Wall Street Journal


THE WALL STREET JOURNAL.

MARKETS | HEARD ON THE STREET

Microsoft and Google Will Both Have t Bear AI's Costs

Artificial-intelligence tools such as ChatGPT could upend cloud computing and search

By Dan Gallagher Follow

Updated Jan. 18, 2023 10:20 am ET

Gift unlocked article | Listen (5 min)

Microsoft CEO Satya Nadella at the World Economic Forum in Davos, Switzerland, on Tuesday.

Slika 6. Članak u Wall Street Journalu o troškovima umjetne inteligencije

Razvoj modela

Kao i svaka računalna aplikacija, i aplikacije pokretane jezičnim modelima prolaze kroz *offline* i *online* fazu razvoja. Online faza, koja nastupa nakon što se model napravi dostupnim dijelu korisnika, pokušava izmjeriti kako korisnici koriste aplikaciju. Ova je faza uobičajena za sve online aplikacije.

Offline faza je ona u kojoj se parametri modela procjenjuju, što u terminologiji strojnog učenja i umjetne inteligencije zovemo *treniranje modela*.

Za postupak treniranja modela trebamo:

- **podatke**
- **model** - odabrati arhitekturu, način inicijalizacije parametara, optimizacijski algoritam...
- **infrastrukturu** - cijeli se postupak obično paralelizira na više procesora u oblaku radi ubrzanja

Podatci

Veliki jezični modeli trebaju ogromne količine podataka. Podatci dolaze iz različitih tekstualnih izvora. Ti izvori dodatno ne moraju biti činjenično točni. Može se raditi o namjerno ili nenamjerno izmišljenoj priči ili zastarjelom podatku. Kvaliteta dostupnih podatka iz tekstualnih izvora će varirati.

Internet tražilice *Google* i *Bing* već dulje vrijeme mogu dati jednostavne odgovore na pitanja. Na Slici 7. dan je primjer upita tko je gradonačelnik grada New Yorka. Na Internetu postoji pregršt informacija o tome, no iskustva rada na tražilicama daju kompanijama koje su ih razvijale bolji uvid u to gdje se kvalitetni podatci nalaze.

Slika 7. Odgovor na pitanje tko je gradonačelnik New Yorka na Bingu

Poznato je da je *Wikipedija* na engleskom jeziku izrazito kvalitetan i ažuran izvor informacija. Brojni ugledni izdavači također imaju pouzdane sadržaje. Dio tih sadr-

žaja javno je dostupan na stranicama tih izdavača, dok je *Wikipedija* enciklopedija na čiji sadržaj nitko ne polaže prava. Izdavači su uložili velik novac u svoj sadržaj, a uvidjeli su da *ChatGPT* i drugi proizvodi temeljeni na velikim jezičnim modelima daju informacije o tom sadržaju, što je dovelo do tužbi poput one koju je pokrenuo *New York Times* protiv *OpenAI* i *Microsofta* oko korištenja sadržaja pod autorskim pravima [3]. S druge strane, neke kompanije sklapaju ugovore za pristup sadržajima izdavača. Nekad su ti ugovori ekskluzivni (kao što je navedeno u [4]) – tehnološka kompanija sklopi s izdavačem ugovor da će isključivo ona imati pristup izdanjima izdavača za potrebe svojih proizvoda.

Modeli

Kako bi čitatelj dobio sliku ideja i izazova modela koji se koriste za velike jezične modele, ovdje ćemo navesti nekoliko primjera i izazova koji oni donose.

n-gram modeli

Ovi su modeli iz rane faze jezičnih modela. Ideja je da se na temelju zadnjih n riječi korisnikovog unosa pokuša predvidjeti ostatak teksta. Ova ideja temelji se na dobro istraženoj teoriji **Markovljevih modela**.

Za dan korisnikov unos $\sigma_1\sigma_2...\sigma_k\sigma_{k+1}...\sigma_{k+n}$ vjerojatnosna razdioba ispisa ovisi samo o zadnjih n riječi unosa. Matematičkim rječnikom, vrijedi:

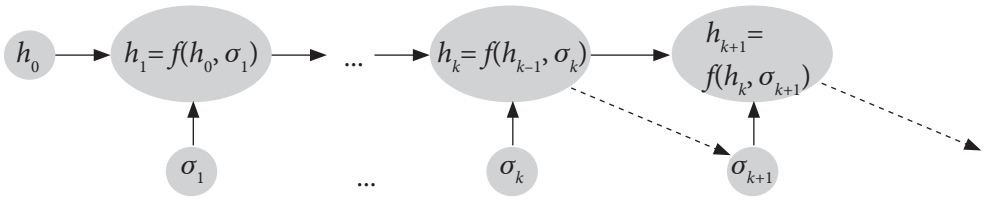
$$\mathbb{P}\left(\cdot \mid \sigma_1\sigma_2...\sigma_k\sigma_{k+1}...\sigma_{k+n}\right) = \mathbb{P}\left(\cdot \mid \sigma_{k+1}...\sigma_{k+n}\right).$$

Dva primjera ilustrirat će nam kako funkcionira ovaj pristup i otkriti njegove nedostatke. Neka je $n = 3$.

- Pretpostavimo da je korisnik unio „U utorak, *nakon kiše bit će i*". Kako se radi o modelu koji gleda zadnji 3-gram, jedini dio unosa koji će se razmatrati bit će „*nakon kiše bit će i*", tako da je vrlo vjerojatno da će jezični model dopuniti ovu rečenicu sa „snijega".
- Mala modifikacija na početku može bitno promijeniti kakav bi ispis trebao biti, što će n-gram modeli zanemariti. Primjerice, unos korisnika poput „Bit će vruće u utorak, *nakon kiše bit će i*" s jednakom vjerojatnošću, kao i u prethodnoj točki, može vratiti „*snijeg*" iako, kad se gleda puna rečenica, to ne bi trebalo biti tako.

Rekurzivni jezični modeli

Ideja ove metode je procesuirati cijeli unos $\sigma_1,...,\sigma_k$ prije nego se proizvede idući simbol. Sve je prikazano na Slici 8. Sve procesuirano sprema se u tzv. skrivenim stanjima – nizu visokodimenzijskih vektora $\left(h_j\right)$ koji se rekurzivno računaju koristeći formulu $h_j = f(h_{j-1}, \sigma_j)$. Idući simbol σ_{k+1} simuliran je na temelju stanja h_k. Izazov je procijeniti funkciju f.

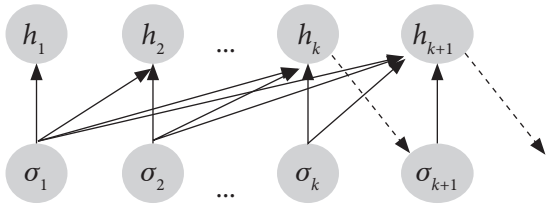


Slika 8. Ideja rekurzivnog jezičnog modela: σ_1, ..., σ_k je unos. Stanja h_1, h_2, ..., h_k su izračunata. Na temelju skrivenog stanja h_{k+1} generira se σ_{k+1} i simboli iza njega.

Ovi su modeli imali poteškoća u održavanju konteksta kroz duže nizove teksta i bili su sporiji zbog sekvencijske prirode obrade podataka. Rekurzivni su se pak modeli suočavali s izazovima poput nemogućnosti dugoročnog pamćenja i paralelizacije[3].

Modeli bazirani na transformerima

Transformeri su riješili ove probleme uvođenjem mehanizma pažnje koji omogućuje modelima da istovremeno obrađuju sve riječi u rečenici i fokusiraju se na relevantne dijelove teksta bez obzira na njihovu udaljenost. Ovo je rezultiralo bržom obradom i boljim razumijevanjem konteksta. Prednosti transformera uključuju brže izvršavanje, učinkovitiju obradu dugih nizova i superiorne performanse u zadatcima poput prevođenja, generiranja teksta i analize osjećaja. Shema, u koju nećemo detaljno ulaziti, dana je na Slici 9. Sa skice se vidi da svako stanje h_k isključivo ovisi o unosu σ_1, ..., σ_k, ali ne i o drugim stanjima h_1, h_2, ..., h_{k-1}.



Slika 9. Arhitektura transformera

Infrastruktura

Procjena parametara složena je operacija. Milijarde dokumenata treba negdje držati, a i obraditi. Potrebna nam je stoga posebna računalna infrastruktura – tzv. **super računalo**. Treniranje traje danima, tjednima, a nekad i mjesecima. Vrlo je izgledno da će neki od servera u podatkovnom centru ispasti iz sustava. Cijelokupni postupak izvodi se u više etapa i ponekad je potrebno pojedinu etapu početi u nekom drugom podatkovnom centru u odnosu na prethodnu. Na Slici 10. ilustrirano je više etapa treniranja *ChatGPT*-a u raznim podatkovnim centrima koje ima Microsoftov računalni oblak *Azure*.

[3]U računarstvu, paralelizacija je postupak u kojem se izvršavanje nekog procesa pokušava podijeliti na više računala kako bi se postupak završio brže i efikasnije. U ovom slučaju, zbog ovisnosti o prethodnim stanjima, kako bismo izračunali stanje h_k, moramo izračunati sva prethodna h_1, h_2, ..., h_{k-1}.

Slika 10. Ilustracija treniranja ChatGPT-a u etapama i podatkovnim centrima diljem svijeta. Ilustracija preuzeta iz [5].

Zahtjevnost infrastrukture potrebne za treniranje velikih jezičnih modela dovodi do pitanja kako uključiti akademsku zajednicu u njihov razvoj. Trenutno malo velikih kompanija s potrebnom infrastrukturom može raditi na njihovom razvoju. Tablica 1. daje nam pregled nekih poznatih modela i nazive tvrtki koje su ih razvile.

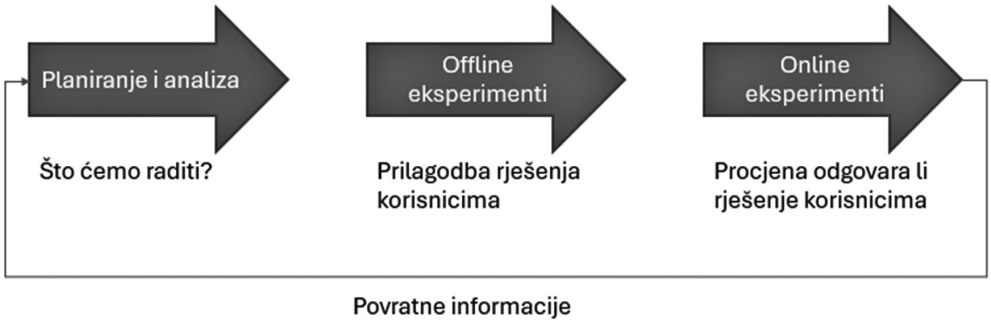
Godina	**Ime modela**	**Broj parametara**	**Kompanija**
2018	BERT	$340 \cdot 10^6$	Google
2019	GPT-2	$1.5 \cdot 10^9$	OpenAI
2020	GPT-3 (GPT-3.5 -> ChatGPT)	$175 \cdot 10^9$	OpenAI
2021	Megatron-Turing NLG	$530 \cdot 10^9$	Microsoft & Nvidia
2022	LaMBDA	$137 \cdot 10^9$	Google
2023	GPT-4 (Bing AI, ChatGPT Plus)	$\approx 10^{12}$	OpenAI
2023	Palm 2 (Google Bard)	$340 \cdot 10^9$	Google

Tablica 1. Pregled nekih popularnih modela razvijenih između 2018. i 2023. godine

Kao što možemo vidjeti iz tablice, većinu modela razvile su velike tehnološke kompanije. Zbog skupoće izrade takvih modela, postavlja se pitanje kako uključiti akademsku zajednicu kako razvoj ne bi ostao isključivo unutar velikih kompanija.

Evaluacija modela

Kod svih modela strojnog učenja i umjetne inteligencije treba napraviti procjenu radi li zadovoljavajuće i bolje od drugog modela ili prethodne verzije.



Offline način

Nakon izgradnje modela, a prije nego što se izloži korisnicima, provodi se temeljito testiranje na skupu podataka koji nije korišten tijekom treniranja. Ovo testiranje obuhvaća procjenu točnosti modela i njegovih performansi. Za ovu svrhu razvijene su specifične tehnike poput *perplexity*, BLEU i ROUGE, koje omogućuju detaljnu analizu i ocjenu kvalitete generiranog teksta.

Online način

Svi modeli moraju zadovoljiti nekoliko uvjeta kako bi korisnici njima bili zadovoljni:

- moraju posluživati milijune korisnika
- moraju raditi 24 sata 7 dana na tjedan
- moraju raditi zadovoljavajućom brzinom

Ovisno o poslovnim ciljevima kompanije koja izrađuje proizvod, kompanije žele poboljšati iduće stvari:

- aktivnost korisnika
- monetizaciju

Više o online eksperimentiranju možete naći u knjizi [5] i članku [6] koji daje njezin pregled.

Mogućnosti i izazovi upotrebe jezičnih modela

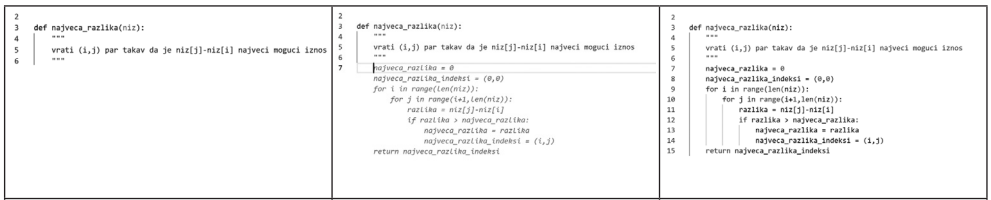
U ovom odjeljku malo ćemo se pozabaviti pitanjem na koje se sve načine pokušavaju koristi veliki jezični modeli.

Sastavljanje teksta

Veliki jezični modeli doista su revolucionarni u načinu na koji pristupaju generiranju i obradi teksta. Oni ne samo da omogućuju sastavljanje različitih vrsta tekstova, poput emailova, blogova i lista, već omogućuju njihovo brzo i učinkovito uređivanje. Osim toga, sposobnost prevođenja teksta na različite jezike čini ih izuzetno korisnima u globaliziranom svijetu. Poboljšanje postojećeg teksta još jedna je značajka koja ističe njihovu adaptabilnost i inteligenciju. Razumijevanje strukture teksta omogućuje im da se nose s kompleksnim zadatcima kao što su formatiranje blogova, pisanje matematičkih izraza pomoću *LaTeX*-a ili kreiranje *HTML* struktura za web-stranice. Sve ove sposobnosti čine ih nezamjenjivim alatom u modernom digitalnom dobu.

Pisanje programskoga koda

Jezični modeli pokrivaju i programske jezike. Od pojave *ChatGPT*-a korisnici su mogli tražiti da im napiše kod u brojnim programskim jezicima. Za potrebe ljudi koji rade u IT industriji, od strane *GitHuba*[4] i *OpenAI*-a razvijen je *GitHub Copilot*–alat za automatsko dovršavanje koda. Ovaj alat korisnicima *Visual Studio Code*, *Visual Studio*, *Neovim* i *JetBrains* integriranih razvojnih okruženja pomaže u automatskom dovršavanju koda. *Copilot* koristi tehnologiju mogućnosti dopunjavanja teksta kako bi predložio sljedeće linije koda ili funkcije temeljene na kontekstu trenutno napisanog koda. Također, *Copilot* može pomoći u generiranju testova, komentara i drugih korisnih segmenata koda, što ga je učinilo poželjnim alatom za programere svih razina iskustva.



Korak 1. *Ime metode u Pythonu i opis*	Korak 2. *GitHub Copilot nudi mogućnost dopune*	Korak 3. *Korisnik pritiskom „Tab“ tipke potvrđuje*

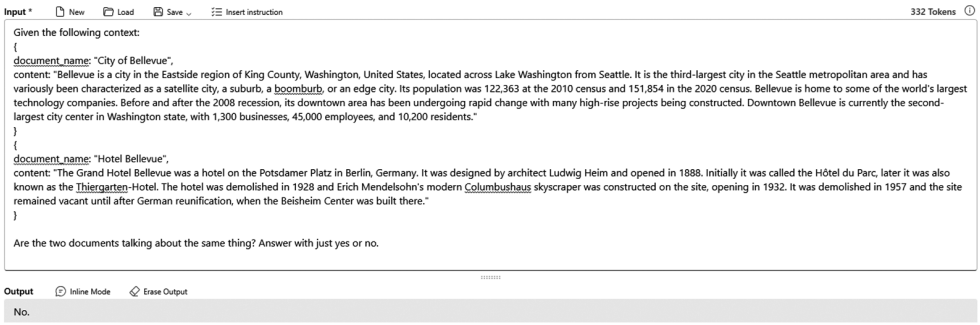
Slika 11. Primjer koraka u korištenju GitHub Copilota

Izrada aplikacija

Velike jezične modele ljudi koriste za različite namjene. Ovdje ćemo pokazati kako se koriste za analizu teksta i izradu aplikacija za komunikaciju s korisnicima.

[4]GitHub je kompanija u vlasništvu *Microsofta*.

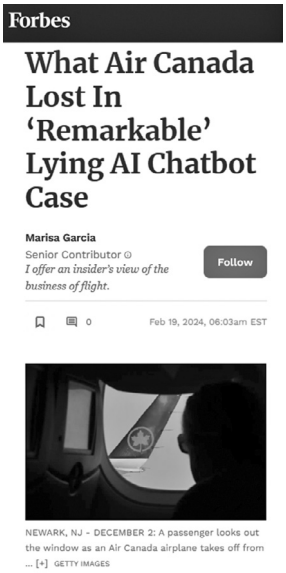
Jezični modeli koriste se za analizu teksta tako što procesuiraju velike količine podataka kako bi razumjeli jezik i kontekst. Ljudi uputama definiraju parametre analize, poput prepoznavanja teme, sentimenta ili strukture teksta. Modeli zatim koriste te upute da procijene tekst i donesu zaključke, bilo da se radi o sažimanju informacija, odgovaranju na pitanja ili generiranju novog teksta na temelju naučenih obrazaca. Na Slici 12. dan je jedan primjer.



Slika 12. Jezični model u ovom slučaju radi usporedbu dvaju dokumenata i donosi zaključak govore li o istoj temi.

Veliki jezični modeli temelj su mnogih suvremenih aplikacija za interakciju s korisnicima, uključujući chat botove. Kada korisnik postavi pitanje ili izrazi zahtjev, jezični model analizira upit i generira odgovor koji oponaša ljudsku komunikaciju. Odgovori koje dobijemo su relevantni i često ih ne možemo razlikovati od onih koje bi dao čovjek. Osim toga, sposobni su učiti iz interakcija, što ih čini sve boljima u razumijevanju nijansi jezika i konteksta tijekom vremena.

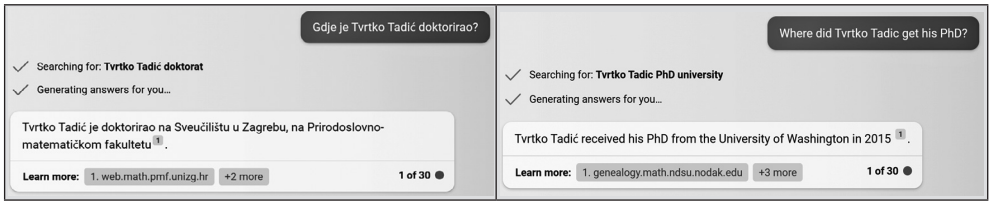
Problem je u oba slučaja što jezični modeli ponekad mogu izbaciti stvari koje nisu točne. Tako je poznat slučaj iz 2023. kad je aplikacija za interakciju s korisnikom kanadske aviokompanije *Air Canada* ponudila popust koji nije postojao. Sudskom odlukom kompanija je morala korisniku priznati taj popust.



Slika 13. Slučaj kad je AirCanada chatbot obećao nepostojeći popust.

Halucinacije

Kao što smo malo prije vidjeli, jezični modeli, koji zapravo predviđaju tekst i simuliraju ga, mogu pogriješiti, a nekada dati i odgovor koji nije u skladu s činjenicama i logikom. Takvu vrstu odgovora zovemo **halucinacijom**. To se može dogoditi zbog nedostatka znanja, konteksta ili razumijevanja.



Slika 14. Bing Chat na hrvatskom tvrdi da je autor doktorirao na Sveučilištu u Zagrebu, dok na engleskom kaže da je to bilo na University of Washington u Seattleu. (Točna je engleska verzija.)

Halucinacije dolaze u oblicima:

- lažnih ili nepostojećih činjenica, imena, datuma, brojeva, događaja
- nekonzistentnog ili proturječnog teksta u odnosu na prethodni kontekst ili znanje
- besmislenog ili nerazumljivog teksta koji nema logike ili strukture
- ...

Ovo su neka autorova osobna iskustva:

- Pozvao polje u strukturi podataka koje nije postojalo: struktura.nepostojece_polje.
- Pri upitu za kontakt čiji e-mail nije imao, model je adresu izmislio.
- Modeli često nisu sigurni radi li se o jednoj ili više osoba.

Granice mogućnosti

Veliki jezični modeli predmet su (usporednih) analiza koje istražuju njihove sposobnosti na različitim jezicima i u različitim zadatcima, kao i studija koje ispituju njihov potencijal u specifičnim industrijama poput zdravstva, obrazovanja itd.

Ovdje ćemo dati nekoliko primjera što su sve ljudi istraživali.

Sposobnosti rješavanja matematičkih zadataka

Model GPT-4 uspio je riješiti pojednostavljeni zadatak s *Međunarodne matematičke olimpijade* održane 2022. godine – IMO 2022. U trenutku kad su autori pripremali izvještaj [8], podatci kojima je model bio izložen bili su prije održavanja olimpijade, a zadatci koji se pojavljuju na IMO, trebaju biti originalni[5], tako da model nije mogao biti izložen ovom zadatku i njegovu rješenju.

[5]Na ovom natjecanju sve zemlje sudionice mogu predložiti zadatak. Prije natjecanja organizatori formiraju listu zadataka koji su ušli u uži odabir, tzv. *shortlist*. Voditelji ekipa iz drugih zemalja pred samo natjecanje trebaju uložiti prigovor ako je zadatak poznat.

GPT-4

Prompt:

Suppose g is a continuous smooth function such that for every $x > 0$, there is **one and only one** $y > 0$ such that $g(x) + g(y) \leq 2xy$. Can you prove that $g(x) = x^2$?

GPT-4:

There are a few possible ways to approach this problem, but one possible method is as follows:

- Assume, for the sake of contradiction, that $g(x)$ is not equal to x^2 for some $x > 0$.
- Then, either $g(x) < x^2$ or $g(x) > x^2$ for that x.
- If $g(x) < x^2$, then let $y = x$. Then $g(x) + g(y) = 2g(x) < 2x^2 = 2xy$.
- If $g(x) + g(y) < 2xy$, then $g(x) + g(y)$ is strictly less than the upper bound of $2xy$ for positive x and y.
- This means that there is some room to increase y slightly and still satisfy the inequality $g(x) + g(y) \leq 2xy$.
- For example, let $y' = y + \epsilon$, where ϵ is a small positive number such that $g(x) + g(y') < 2xy$. This is possible because g is continuous and smooth, so it does not have any jumps or kinks that would prevent such a choice of epsilon.
- Then, y' is another positive number such that $g(x) + g(y') \leq 2xy'$, which contradicts the assumption that y is the one and only one positive number that satisfies this condition given x.
- Therefore, if $g(x) < x^2$, then $g(x) + g(y)$ cannot be strictly less than $2xy$, and it must be equal to $2xy$.
- If $g(x) > x^2$, then let y^* be the unique positive solution of $g(x) + g(y) = 2xy$. Then $g(y^*) < (y^*)^2$, since otherwise we would have $g(x) + g(y^*) > 2xy$, which contradicts the existence of y^* given x.
- But then, by the same argument as before, we can let $x = y = y^*$ and obtain a contradiction, since $g(y) + g(y) = 2g(y) < 2y^2 = 2yy$, which contradicts the uniqueness of y given y.
- Therefore, $g(x)$ must be equal to x^2 for all $x > 0$, as required.

Slika 15. Pojednostavljeni zadatak s IMO 2022. i rješenje preuzeto iz [8].

Premda je uspio riješiti ovaj zadatak, autori izvještaja pokazuju da ovaj model često radi najjednostavnije greške.

Tip greške	Postotak
aritmetička greška (uključuje i prebrojavanje)	68 %
krivo razumijevanje problema	10 %
krivi pristup	22 %

Zanimljiv je primjer s polinomom u kojem je GPT-4 bio uspješan u 75.2 % slučajeva sa Slike 16. GPT-4 dobio je tri različite nul-točke slučajno odabrane iz skupa $\{-10,-9,...,9,10\}$ polinoma p stupnja 3, kao i vrijednost $p(1)$ također odabranu slučajno iz istog skupa. Pitanje je bilo koja je vrijednost od $|p(0)|$?

GPT-4 memorization test by alternating the numbers

Prompt: `If a degree 3 polynomial satisifies p(x) = 0 for x = -3, 8, 5 and p(1) = 10, what is |p(0)|?`

Slika 16. Permutacija brojeva dovodi do toga da uglavnom dobijemo točan odgovor, ali ne uvijek.

Jezični modeli uglavnom proizvode tekst i nemaju mogućnosti naprednog računanja. Ali imajući u vidu mogućnosti sustava za računalnu algebru (poput *Mathematica*, *Maple*, *SAGE*, ...), nije neobično da će ovakve stvari računalo u budućnosti moći

riješiti. *Google* je nedavno objavio da je njihov sustav umjetne inteligencije riješio čak četiri od šest zadataka s *Međunarodne matematičke olimpijade* održane 2024. godine [1], no to još nije javno dostupan proizvod.

Programerske sposobnosti

GPT-4 zna pisati, analizirati i simulirati kod. Prošao je simulaciju intervjua. Na standardnim testovima postiže rezultate slične ljudima. U Tablici 2. vidimo rezultate raznih modela (preuzeto iz [8]) kad su im dani problemi koji se javljaju na intervjuima za posao u velikim IT kompanijama:

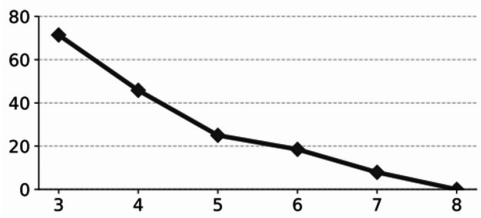
Zadatci	Lagani		Osrednji		Teški		Sveukupno	
Broj pokušaja *K* Model	*K = 1*	*K = 5*	*K = 1*	*K = 5*	*K = 1*	*K = 5*	*K = 1*	*K = 5*
GPT-4	68.2	86.4	40.0	60.0	10.7	14.3	38.0	53.0
Text-davinci-003	50.0	81.8	16.0	34-0	0.0	3.6	19.0	36.0
Code-davinci-002	27.3	50.0	12.0	22.0	3.6	3.6	13.0	23.0
Ljudi (LeetCode korisnici)	72.2		37.7		7.0		38.2	

Tablica 2. Postotci riješenosti zadataka koji se pojavljuju na intervjuima u IT kompanijama

Treba napomenuti kako su na *GitHub* repozitoriju javno dostupne ogromne količine programerskog koda, što je omogućilo da jezični modeli imaju mogućnost stvaranja novog koda. Mogućnosti su i dalje ograničene za brojne složenije probleme.

Analiza grafova

U računarstvu i matematici analiza grafa (mreže) jedan je od većih izazova. U radu [2] istražene su sposobnosti velikih jezičnih modela u razumijevanju veza u grafovima. U nezanemarivom broju slučajeva za jednostavne grafove mogu naći odgovore, no ta mogućnost opada za složenije grafove s više vrhova. Kao što se vidi na Slici 17., traženje najkraćeg puta za 8 i više vrhova postaje nemoguće za velike jezične modele.



Slika 17. Traženje najkraćeg puta u grafu.
Postotak uspješnosti pada što je put dulji.

Fizika umjetne inteligencije

Program „Physics of AI" koji se razvija u *Microsoft Researchu* usmjeren je na razumijevanje dubokog učenja kroz *ideje iz fizike*. Ovaj pristup uključuje istraživanje fenomena putem kontroliranih eksperimenata i izgradnju teorija pomoću pojednostavljenih matematičkih modela. Cilj je bolje razumijevanje i unapređenje inteligencije u velikim jezičnim modelima. Trenutno se vrlo malo razumije zašto modeli putem transformatora uspješno izvršavaju zadatke. Više o svemu možete doznati u javno dostupnom predavanju [9] objavljenom na *YouTubeu*.

Odgovoran pristup umjetnoj inteligenciji

Odgovorna umjetna inteligencija (RAI) odnosi se na razvoj i implementaciju AI sustava koji su transparentni, nepristrani, odgovorni i koji slijede etičke smjernice. To je posebno važno jer AI sustavi postaju sveprisutni u mnogim aspektima naših života, od zdravstva do prijevoza. Etički izazovi uključuju osiguravanje pravednosti i izbjegavanje pristranosti, što znači da AI sustavi trebaju tretirati sve korisnike jednako, bez obzira na njihovu pozadinu. Također, postoji potreba za odgovornošću, gdje bi pojedinci i organizacije trebali biti odgovorni za odluke koje donose njihovi AI sustavi. Transparentnost je ključna kako bi korisnici mogli razumjeti odluke koje AI donosi, a privatnost je temeljna kako bi se zaštitili osobni podatci korisnika. Jezični modeli, kao što su oni koji omogućuju razumijevanje i generiranje prirodnog jezika, donose dodatne izazove jer se moraju nositi s nijansama jezika i kulture te potencijalno utjecati na mišljenja i odluke ljudi. Mnoge kompanije definirale su svoje principe u javno dostupnim dokumentima poput [10].

Zaključak

Jezik je temeljno sredstvo komunikacije, a jezični su modeli ključni u razumijevanju i generiranju prirodnog jezika. Od predviđanja sljedeće riječi do pisanja punih tekstova – jezični su se modeli vrlo brzo razvili, a postali su sposobni i raditi neke stvari koje su do sada mogli raditi samo ljudi, kao što je rješavanje zadatka iz programiranja. Također, jezični modeli omogućili su korisnicima komunikaciju s računalom na prirodnom jeziku i time iznimno povećali korisnost tehnologije.

Oni zahtijevaju složenu infrastrukturu i ogromne količine podataka pa postupak njihova stvaranja traje jako dugo. Ipak, unatoč impresivnim rezultatima, i dalje rade greške odnosno halucinacije. Dok tehnologija napreduje, možemo očekivati da će jezični modeli ubrzo postati još sofisticiraniji, otvarajući nove mogućnosti za razumijevanje i interakciju s prirodnim jezikom.

Literatura:

1. Google DeepMind, „AI achieves silver-medal standard solving International Mathematical Olympiad problems", https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/ (25. 7. 2024.)
2. H. Wang, S. Feng, T. He, Z. Tan, X. Han and Y. Tsvetkov, „Can Language Models Solve Graph Problems in Natural Language?", https://arxiv.org/abs/2305.10037. (srpanj 2024.)
3. M. M. Grynbaum and R. Mac", The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work," *The New Zork Times,* (27 12 2023.)
4. L. Jones, „Google Gains Exclusive Access to Reddit Data for AI, Restricting Rival Search Engines", WinBuzzer, https://winbuzzer.com/2024/07/25/google-gains-exclusive-access-to-reddit-data-for-ai-xcxwbn/. (26. 7. 2024.)
5. Microsoft, „What runs ChatGPT? Inside Microsoft's AI supercomputer | Featuring Mark Russinovich", https://www.youtube.com/watch?v=Rk3nTUfRZmo&t=12s. (27. 7. 2024.)
6. R. Kohavi, D. Tang and Y. Xu, Trustworthy Online Controlled Experiments : A Practical Guide to A/B Testing, Cambridge University Press, 2020.
7. T. Tadić, „Online eksperimenti - iskustva velikih kompanija", *Poučak,* 70–77, 2022.
8. S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro and Y. Zhang, „Sparks of Artificial General Intelligence: Early experiments with GPT-4", https://arxiv.org/abs/2303.12712. (srpanj 2024.)
9. S. Bubeck, „Physics of AI", https://www.youtube.com/watch?v=XLNmgviQHPA. (srpanj 2024.)
10. Microsoft , „Microsoft Responsible AI Standard v2," 7 2022. [Online]. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf (srpanj 2024.)
11. OpenAI, „GPT-4 Technical Report", https://arxiv.org/abs/2303.08774. (srpanj 2024.).
12. R. Cotterell, A. Svete, C. Meister, T. Liu and L. Du, „Formal Aspects of Language Modeling", https://arxiv.org/abs/2311.04329. (srpanj 2024.)
13. T. Tadić and M. Smiljanić, „Veliki jezični modeli – Što se sve promijenilo razvojem ChatGPT-a?", https://youtu.be/LZPK34g76No. (27. 7. 2024.)