Veliki jezični modeli – potraga za umjetnom inteligencijom

Reni Banov[1]

Unazad deset godina zamjećujemo veliki porast zanimanja za rješenjima ostvarenim primjenom umjetne inteligencije na konkretnim problemima iz različitih područja, pa se ona, primjerice, koristi za e-trgovine na međumrežju, za obrazovne aktivnosti, za autonomnu navigaciju, za upravljanje u robotici, te u brojnim drugim područjima. Međutim, niti jedno od brojnih rješenja nije privuklo medijsku pozornost kao nedavna primjena umjetne inteligencije za generiranje i „razumijevanje" tekstualnih zapisa prirodnih jezika [1]. Takvi se modeli u pravilu nazivaju velikim jezičnim modelima (eng. *Large Language Models*, skraćeno LLM) jer su pripremljeni procesom učenja neuronskih mreža na velikim skupovima ulaznih podataka, prvenstveno iz tekstualnih zapisa (novinski i znanstveni članci, mrežne stranice, komunikacija na forumima itd.) s međumrežja. Iako je primjena takvih modela prividno neograničena, ona u brojnim aspektima ljudske komunikacije pokazuju manjkavosti. U ovom članku pokušat ćemo razjasniti osnovne matematičke pojmove povezane s tim modelima, te djelomično prikazati pozitivne i negative strane njihove primjene. Nastavno na ovaj, u sljedećem ćemo članku na jednostavnom primjeru primjene jezičnog modela u obrazovnom procesu (pisanje eseja) prikazati izvjesne nedostatke bez dublje analize metodoloških posljedica na sam proces obrazovanja. Kako bismo razumjeli dobivene rezultate, važno je prije same primjene jezičnog modela upoznati se s njima, kao i s tehnološkim napretkom koji je doveo do intenziviranja njihove primjene.

Uvod

Dohvat velikog opsega podataka koji se generira u različitim sferama ljudske aktivnosti dovodi do potrebe za efikasnim nalaženjem njihove smislene interpretacije. Jedan od pristupa tom problemu zasniva se na prikupljanju što veće količine relevantnih podataka za određeni problem te primjenu umjetne inteligencije za otkrivanje „smislene" interpretacije. Sposobnost nalaženja vjerodostojne interpretacije imanentna je kompetentnim osobama koje s dovoljnom razinom znanja i kritičkim razmišljanjem pronalaze zakonitosti u tim podatcima, što su osobine svojstvene čovjeku.

[1]Reni Banov, Tehničko veleučilište u Zagrebu

U novije vrijeme zamjetan je pomak u pristupu koji se u svojoj osnovi zasniva na pitanju kako iz velikog broja prikupljenih podataka uspostaviti pouzdan inteligentan sustav neposrednom primjenom matematičkih metoda linearne algebre [2]. Pristup nalaženja smislene interpretacije promatra se kao problem prikupljanja dovoljne količine podataka te njihove efikasne računalne obrade umjetnim neuronskim mrežama.

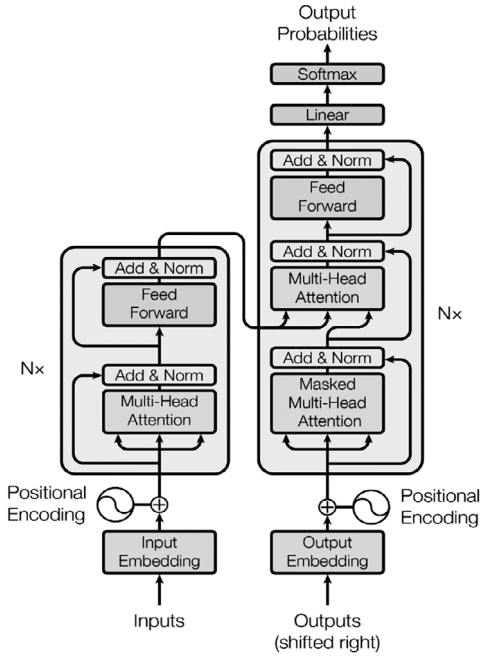
Neuronske mreže zasnivaju se na principu oponašanja funkcije biološkog neurona pomoću višedimenzionalnih regresijskih modela (linearnih ili nelinearnih) te na izgradnji veza između slojeva umjetnih neurona kojima se simulira funkcionalna sinaptička povezanost prisutna u ljudskom mozgu [3]. Kod neuronskih mreža uočavamo trend kako se postupkom učenja na sve većim količinama ulaznih podataka postižu sve bolji rezultati u fazi njihove primjene na konkretnim problemima. Primjerice, široko su rasprostranjene neuronske mreže koje postižu iznimnu točnost u domeni prepoznavanja simbola, kao što su registarske pločice na cestovnim vozilima. Uz prisutnost navedenih činjenica te ozbiljnim porastom računalne snage, osobito pojavom grafičkih procesora sposobnih za efikasno istovremeno (paralelno) obavljanje aritmetičkih operacija na više podataka, razvijaju se „inteligentni" sustavi zasnovani na velikim neuronskim mrežama. Obično se veličina neuronske mreže povezuje s brojem umjetnih neurona te brojem parametara (težina) koji određuju jačinu njihove međusobne veze. Procjenjuje se kako ChatGPT-4 neuronska mreža brojem neurona i veza između njih dostiže broj bioloških neurona (reda 10^{11}) te se približava broju sinaptičkih veza (reda 10^{15}) u ljudskom mozgu.

Osim potrebne računalne snage za izgradnju neuronske mreže, veliki jezični modeli zahtijevaju iznimnu količinu ulaznih podataka u postupku nadziranog i ne nadziranog učenja. Za ilustraciju veličine samog projekta izgradnje takvog sustava možemo razmotriti statističke parametre tijekom pripreme i izvođenja. ChatGPT sustav izgrađen je pomoću sadržaja međumrežnih stranica iz različitih izvora (Common Crawl, WebText2, Books1, Books2, ...), od kojih gotovo dvije trećine ulaznih podataka čini Common Crawl arhiva koja u ovom trenutku (travanj 2025.) sadrži približno 2,75 milijarde mrežnih stranica ukupne veličine oko 455 TB nefiltriranih podataka [4]. Međutim, za potrebe učenja neuronske mreže ChatGPT sustava iz navedenih arhiva odabrano je približno 300 milijardi riječi (cca. 0,6 TB komprimirano), a prema nekim izvorima [5] troškovi samog procesa učenja procijenjeni su na više od 12 milijuna US dolara, dok procjene za dnevne operativne troškove dostižu i do 700 tisuća US dolara. Svi ti podatci ukazuju na izniman opseg projekta te na veličinu problema koji se pokušava riješiti tim pristupom.

Osnovni elementi LLM modela

Prije nego što primijenimo ChatGPT za izradu jednog eseja upoznajmo arhitekturu velikih jezičnih modela koja predstavlja „mozak" tog sustava. Ideja velikih jezičnih modela svoj početak nalazi u radu nekolicine inženjera iz tvrtke Google [6] koji

su razvili model neuronske mreže za prevođenje tekstova s engleskog jezika na njemački jezik. Njihova ideja zasniva se na specijaliziranoj arhitekturi koja koristi pojam pozornosti (eng. *attention*) za predviđanje sadržaja niza znakova (eng. *token*) tijekom postupka prevođenja teksta. Arhitekturu nazivaju pretvorbenom (eng. *transformer*) te je shematski prikazana na sljedećem dijagramu preuzetom iz njihova rada.



Slika 1. Transformer arhitektura za LLM model (izvor [6])

Iz dijagrama je odmah razvidno kako se radi o složenom modelu koji nije samo velika neuronska mreža nastala spajanjem brojnih međuslojeva, već se zasniva na dva niza pretvorbenih blokova (oznaka Nx na dijagramu predstavlja broj ponavljanja takvih blokova, obično 6 ili više) za tekstualni ulaz (vektor $(\boldsymbol{x}_1,...,\boldsymbol{x}_n)$) te mogući slijedni tekstualni nastavak (vektor $(\boldsymbol{y}_1,...,\boldsymbol{y}_n)$) iz ulaznog konteksta. Naime, uočava se višeslojna arhitektura pretvorbenog bloka koja se u osnovi sastoji od dijela koji izračunava vrijednost za pozornost niza znakova (*Multi-Head Attention* dio), te dijela koji pomoću ocijenjene pozornosti smješta niz znakova na različitim pozicijama (*Feed Forward* dio) u odabranom kontekstu. Ideja pretvorbenog bloka dolazi iz intuitivnog razmišljanja kako se višeslojnom arhitekturom može preciznije ocijeniti „značenje" ulaznih riječi (tokena) za odabrani jezični kontekst. Takav pristup izgradnje jezičnih konstrukcija zasniva se na međusobnom povezivanju riječi iz prethodnih i budućih slojeva sa susjednim riječima iz sloja na kojem se računa ocjena pozornosti. Znamo kako semantika neke riječi u određenom kontekstu može biti povezana s brojnim riječima koje ne moraju biti u neposrednoj blizini unutar jedne jezične konstrukcije. Stoga je za ocjenu pozornosti nužno odrediti model kojim se može ocijeniti

„vrijednost" neke riječi u odnosu na ostatak teksta iz konteksta. Model za određivanje pozornosti je upravo takav model koji omogućava određivanje vrijednosti za riječ u širem kontekstu jezične konstrukcije. U osnovi pojam konteksta može se primijeniti na dva načina u modelu za određivanje pozornosti. Jedan način (pristup) za određivanje pozornosti zasniva se na korištenju svih riječi iz prethodnog sloja (ili slojeva), dok se drugi način (dvosmjerni) ne zasniva samo na korištenju riječi iz prethodnih slojeva, već i na korištenju riječi koje predstavljaju moguće (najčešće generirane) jezične konstrukcije.

Određivanje pozornosti

Prvo pitanje koje se postavlja u oba pristupa jest određivanje pozornosti jedne riječi $\boldsymbol{x}_i$ spram druge riječi $\boldsymbol{x}_j$ to jest, kako odrediti numeričku vrijednost $\alpha_{i,j}$ za usporedbu riječi iz konteksta. Budući da su riječi predstavljene vektorima, možda je najjednostavniji način određivanja vrijednosti onaj pomoću ujednačenog (normiranog) koeficijenta proporcionalnosti iz skalarnog umnoška

$$\alpha_{ij} = \frac{\sigma(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)}{\sum_k \sigma(\boldsymbol{x}_i \cdot \boldsymbol{x}_k)}, \quad j \leq i \tag{1}$$

gdje se za funkciju σ uobičajeno koristi neka od aktivacijskih funkcija za neuronske mreže, najčešće normirana eksponencijalna (sigmoidna) funkcija. Na taj se način dobivene vrijednosti (težine) koriste za generiranje jednostavnog vektora mogućih slijednih riječi za odabranu riječ $\boldsymbol{x}_i$

$$a_i = \sum_{j \leq i} \alpha_{ij} \boldsymbol{x}_j . \tag{2}$$

Međutim, pretvorbeni se blok ne zadržava samo na tom jednostavnom modelu mogućih slijednih riječi, već se primjenjuje sličan prošireni model za određivanje doprinosa riječi u širem kontekstu. Prošireni model u pretvorbenom bloku određuje pozornost pojedine riječi u funkciji njezine uloge u širem kontekstu kroz tri aspekta:

1. kao odabrani *fokus pozornosti za sve prethodne ulazne tekstove* pretvorbenog bloka, koji se u širem kontekstu naziva ***upit*** (eng. *query*),
2. kao uloga *prethodnog ulaza za trenutni fokus pozornosti* pretvorbenog bloka, gdje se naziva ***ključem*** (eng. *key*),
3. te kao njezina ***vrijednost*** (eng. *value*) za *računanje slijednih nastavaka trenutnog fokusa* iz pretvorbenog bloka.

Za opisivanje svakog od navedenih aspekata u pretvorbenom modelu uvode se tri težinske matrice koje imaju funkciju projekcije ulaznog vektora $\boldsymbol{x}_i$ za odabranu riječ na njezinu reprezentaciju u ulozi upita, ključa ili vrijednosti. Drugim riječima, pomoću tih matrica provodi se transformacija vektora $\boldsymbol{x}_i$ u nove vektore za uloge upita ($\boldsymbol{q}_i$), ključa ($\boldsymbol{k}_i$) i vrijednosti ($\boldsymbol{v}_i$),

$$\boldsymbol{q}_i = \boldsymbol{x}_i \boldsymbol{W}^q, \qquad \boldsymbol{k}_i = \boldsymbol{x}_i \boldsymbol{W}^k, \qquad \boldsymbol{v}_i = \boldsymbol{x}_i \boldsymbol{W}^v, \tag{3}$$

te se pomoću njih određuje slijedni vektor u pretvorbenom bloku. Tri se težinske matrice izračunavaju tijekom faze učenja neuronske mreže za pretvorbene blokove, pa je za njihovo određivanje iznimno važna konvergencija metode kojom se ažuriraju elementi u tim matricama. Najčešće korištena metoda za ažuriranje elemenata težinskih matrica u neuronskim mrežama zasniva se na primjeni gradijentne metode za određivanje lokalnog minimuma višedimenzionalnih funkcija [7].

Zapažamo kako računanje elemenata tih matrica u osnovi predstavlja operaciju skalarnog umnoška dvaju vektora koji, naravno, može rezultirati velikim pozitivnim ili negativnim vrijednostima. Stoga primjena funkcije σ (u biti eksponencijalne funkcije) na tako izračunatim vrijednostima može dovesti do gubitka numeričke točnosti te posljedično do gubitka točnosti gradijenta tijekom faze učenja. Da bi se izbjegao taj neželjeni gubitak točnosti gradijenta, vrijednost skalarnog umnoška normira se dijeljenjem s dimenzijom (brojem elemenata) vektora $\boldsymbol{q}_i$, $\boldsymbol{k}_i$. Kako se radi o vektorima istih dimenzija (d_h) unutar jednog pretvorbenog bloka, konačan oblik proširenog modela za generiranje mogućih slijednih riječi definira se pomoću tri vektora $\boldsymbol{q}_i$, $\boldsymbol{k}_i$, $\boldsymbol{v}_i$ (3) na sljedeći način; najprije računanjem normiranog skalarnog umnoška za ulogu $\boldsymbol{x}_i$ riječi za upit, i ulogu $\boldsymbol{x}_j$ riječi za ključ

$$\beta_{ij} = \frac{\boldsymbol{q}_i \cdot \boldsymbol{k}_j}{\sqrt{d_h}}, \tag{4}$$

te određivanjem slijednog vektora $\boldsymbol{a}_i$ (2) pomoću koeficijenata proporcionalnosti

$$\alpha_{ij} = \frac{e^{\beta_{ij}}}{\sum e^{\beta_{ij}}}, \qquad j \le i\,. \tag{5}$$

Napomenimo kako koeficijent proporcionalnosti određen izrazom (5) ustvari predstavlja funkciju distribucije (funkcija *softmax*) pojavljivanja riječi u slijednom vektoru iz konteksta mogućih tekstualnih nastavaka.

Prikazani postupak određivanja numeričke vrijednosti za pozornost proveden je za jedan korak na odabranoj riječi, te se može jednostavno zapisati u obliku pogodnom za paralelno izvođenje na modernim grafičkim procesorima. Naime, ukoliko ulazne riječi zapišemo u obliku matrice

$$\boldsymbol{X} = \left[\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_N \right], \tag{6}$$

gdje je N broj riječi (kontekstna duljina, npr. 4000 riječi) koje obrađujemo, tada se postupak može zapisati u matričnom obliku

$$\boldsymbol{A_I} = Pozornost(\boldsymbol{X}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_h}}\right)\boldsymbol{V}\,, \tag{7}$$

gdje su matrice za računanje pozornosti određene izrazima

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^q, \qquad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^k, \qquad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^v \tag{8}$$

Također je važno napomenuti kako se umnožak $\boldsymbol{QK}^T$ u izrazu računa za sve kombinacije riječi u ulozi upita i ulozi ključa, što nije potrebno za one riječi za koje znamo kako slijede iz upita. Stoga se za potrebe optimizacije izračuna umnoška, matrice obično zapisuju u trokutastom obliku za koji se može efikasno paralelizirati izvođenje operacije množenja na grafičkim procesorima.

Višeslojna pozornost

Drugo pitanje koje se pojavljuje prilikom korištenja LLM modela je kako takav model ispravno formulira odgovore u kontekstu, to jest kako model istovremeno razlikuje odnos između riječi u funkciji semantičke važnosti, gramatičke točnosti, sintaktičke ispravnosti, te ostalih parametara iz komunikacije prirodnim jezicima. Model pozornosti iz prethodnog dijela definiran je s tri matrice težina $\boldsymbol{W}^q$, $\boldsymbol{W}^k$, $\boldsymbol{W}^v$ koje se obično formiraju tijekom faze učenja LLM modela na selektiranim podatcima spram odabranog jezičnog aspekta koji opisuju (semantika, gramatika, sintaksa, ...), drugim riječima, u LLM modelu formiraju se skupovi tih matrica koji definiraju pretvorbeni blok za određeni jezični aspekt. Takva organizacija omogućava paralelnu obradu ulaznog teksta prema specifičnom aspektu te generiranje izlaznog teksta koji najbolje odgovara ulaznom prema pravilima koji definiraju odabrani aspekt jezika. Pretvorbeni blokovi za specifične aspekte jezika nazivaju se „glavama" (eng. *head*) u LLM modelu i služe za ocjenu pozornosti prema kriterijima odabranog aspekta.

Poznavanjem pozornosti za pojedini aspekt jezika ($\boldsymbol{A}_l$ iz formule 7) možemo ukupnu vrijednost pozornosti odrediti primjenom jednostavne formule

$$\boldsymbol{A} = \mathit{VišeslojnaPozornost}(\boldsymbol{X}) = \left(\boldsymbol{A}_1 \oplus \boldsymbol{A}_2 \oplus \cdots \oplus \boldsymbol{A}_h\right)\boldsymbol{W}^o, \tag{9}$$

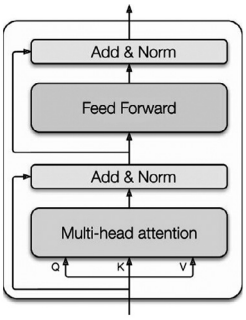
gdje se XOR kombinacija pozornosti za jezične aspekte množi s matricom težina ($\boldsymbol{W}^o$) za izlazni sloj neuronske mreže pretvorbenog bloka. Novije implementacije jezičnih modela, matricu težina za izlazni sloj, personaliziraju za korisnika sustava, to jest težine u matrici dodatno su prilagođene specifičnim zahtjevima (preferencijama) korisnika.

Pomoćni slojevi

Premda određivanje pozornosti predstavlja osnovnu funkciju pretvorbenog bloka, za njegovu potpunu funkcionalnost potrebni su dodatni slojevi neuronskih mreža (blokovi): za transformaciju slijednih vektora (eng. *feedforward layers*), te za neposredno povezivanje (eng. *residual layers*) i normalizaciju izračunatih vrijednosti (eng. *normalizing layers*). Funkcija transformacije slijednih vektora izvedena je pomoću dvoslojnih (jedan skriveni sloj) Perceptron neuronskih mreža [7] te se njome provodi linearna transformacija slijednog vektora određena izrazom

$$f(\boldsymbol{x}) = \max\left(0, \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1\right)\boldsymbol{W}_2 + \boldsymbol{b}_2, \tag{10}$$

gdje su $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ dvije matrice težina, a $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ dva vektora posmaka. Težine za Perceptron neuronske mreže određene su u fazi treniranja jezičnog modela te su specifične za poziciju na kojoj se primjenjuje transformacija. Slojevi za neposredno povezivanje koriste se za prijenos stanja neurona (informacija) iz nižih (najčešće ulaznih) prema nivoima bližim izlaznim neuronima, te se često koriste za poboljšanje brzine i točnosti učenja primjenom gradijentne metode. U pretvorbenom bloku slojevi za neposredno povezivanje koriste se u jednostavnom obliku (bez težina) za prijenos stanja ulaznih vektora do slijednih vektora na koje se primjenjuje postupak normalizacije izračunatih vrijednosti.



Slika 2. Arhitektura pretvorbenog bloka (Izvor [8])

Ulazni se vektori zbrajaju sa slijednim vektorima iz bloka za transformaciju ili sa slijednim vektorima iz bloka za određivanje pozornosti te se nakon normalizacije koriste za ulazne vektore u narednim pretvorbenim blokovima. Brojni su načini normalizacije slijednih vektora koji se primjenjuju u dubokim neuronskim mrežama te se primarno upotrebljavaju za poboljšanje performansi neuronske mreže tijekom učenja. Kako su slijedni vektori $\boldsymbol{x}$ u osnovi slučajni vektori iz višedimenzionalne normalne razdiobe, primjenjuje se postupak njihove standardizacije, to jest transformacija

$$\boldsymbol{z} = \frac{\boldsymbol{x} - \mu}{\sigma}, \tag{11}$$

gdje su vrijednosti parametara μ, σ zasebno određene za svaki slijedni vektor (n je dimenzija vektora)

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2}\,. \tag{12}$$

Konačna funkcija sloja normalizacije određena je jednoslojnim Perceptron modelom neuronske mreže, gdje se normalizirani slijedni vektor $\boldsymbol{z}$ transformira linearnom funkcijom

$$y = \alpha z + \beta \tag{13}$$

određenom s dva parametra. Parametri α, β određuju se tijekom faze učenja za svaki sloj normalizacije.

Izazovi kod jezičnih modela

Premda su posljednjih godina veliki jezični modeli značajno uznapredovali, prvenstveno zahvaljujući tehnološkim dostignućima grafičkih procesora, ipak će njihova upotreba postaviti brojne izazove pred potencijalne korisnike i pružatelje usluga njihova korištenja. U osnovi izazove možemo podijeliti u nekoliko cjelina [9]:

- sociološki, zbog utjecaja na društveno okruženje gdje se intenzivira korištenje: sigurnost, privatnost, dugotrajna ovisnost, nejednakost, regulatorni aspekt, etičke posljedice;
- fundamentalni, kao posljedica primjene umjetnih neuronskih mreža: pretreniranost mreže, zaboravljivost, halucinacije, ograničenost;
- tehnološki, koji proizlaze iz tehnologije primijenjene za implementaciju arhitekture jezičnih modela: komercijalni troškovi, konkurentnost, održivost, rad u stvarnom vremenu;
- znanstveni, uzrokovani nedovoljno razrađenom teorijskom podlogom modela: interpretacija i razumijevanje dobivenih rezultata, generalizacija za primjenu na drugim područjima, i brojni drugi.

Međutim, zbog privlačnosti korištenja možda najveći izazov za velike jezične modele tek slijedi nakon njihove masovne primjene u svakodnevnom radu. Od samih početaka primjene neuronskih mreža poznata je činjenica kako neuronske mreže i modele zasnovane na njima nije uputno učiti (trenirati) na podatcima koje mreža sama generira. Naime, neželjena posljedica takvog pristupa je degenerativni proces u neuronskoj mreži koji dovodi do pada njezine funkcionalnosti [10]. Kako je primjena velikih jezičnih modela sve popularnija, u doglednoj budućnosti možemo očekivati povećanje broja javno dostupnih dokumenata generiranih (ili korigiranih) pomoću LLM modela. Budući je izgradnja velikih jezičnih modela zahtjevan postupak, osobito u fazi učenja, pružatelji usluge korištenja suočit će se s pitanjima poput: kako i gdje pronaći dovoljne količine originalnih dokumenata za nastavak razvoja velikih jezičnih modela, kao i koji su dometi modela u trenutnom stadiju razvoja. Za pretpostaviti je kako će pružatelji usluga po tim pitanjima tražiti pojačani nadzor nad sadržajima koje generiraju veliki jezični modeli, premda nije jasno na koji bi se način takav nadzor uopće mogao provoditi nad postojećim i budućim dokumentima na međumrežju, s obzirom na razne čimbenike poput njihove brojnosti, različitih formata i dr.

Komentar na kraju

Alan Turing se davne 1950. godine u svom poticajnom radu [11] zapitao mogu li računala (strojevi) jednoga dana dostići razinu ljudske inteligencije ili je čak nadmašiti? U tom radu postavljeno je pitanje: možemo li od računala očekivati sposobnost intelektualnog napretka u smislu samostalnosti kod učenja i vlastitog razvoja te u konačnici svjesnog razmišljanja. Turing je već tada naslućivao kako će razvoj računala

s vremenom dovesti do njihove primjene na zadatcima gdje se traže tipično ljudske osobine: učenje, razmišljanje i prilagođavanje nepoznatim situacijama. Proteklo je 75 godina od Turingova rada, računala su nemjerljivo snažnija u usporedbi s računalima toga doba, pa je i Turingova vizija računala koja imitiraju čovjekove misaone procese sve bliža. Premda još uvijek nije u cijelosti ostvarena, kontinuirani napredak u razvoju velikih jezičnih modela smanjuje razliku između ljudskog razmišljanja i njegove imitacije barem u segmentu komunikacije prirodnim jezicima, dok svjesnost i nadalje ostaje nedohvatljiva računalima. Neki od najvećih mislilaca današnjice smatraju kako svjesnost u konačnici nije izračunljiva prema Turingovoj definiciji [12], drugim riječima, kako ipak nije moguće izraditi računalni program (Turingov stroj) koji se ponaša svjesno. Možda je i sam Turing dao odgovor na to pitanje u svom radu riječima: *The only way by which one could be sure that a machine thinks is to be the machine and to feel oneself thinking*.

Literatura:

1. T. Tadić, *Jezični modeli*, Poučak br. 99, 2024, str. 28-42.
2. G. Strang: „Linear Algebra and Learning from Data," Wellesley Cambridge Press, 2019.
3. R. Banov, A. Valent, J. Anušić, *Neuronske mreže za početnike*, Poučak br. 90, 2022, str. 24-34.
4. https://commoncrawl.org/blog/march-2025-crawl-archive-now-available (14. 4. 2025).
5. Više autora, *Compressing Large-Scale Transformer-Based Models: A Case Study on BERT* , Transaction of the Association for Computational Linguistics 2021; Vol. 9; 1061-1080. doi:10.1162/tacl_a_00413
6. Više autora, *Attention is All you Need*, Advances In Neural Information Processing Systems, Curran Associates, Inc., Vol. 30, 2017.
7. C. C. Aggarwal, „Neural Networks and Deep Learning," Springer, 2018.
8. V. Lialin, V. Deshapande, X. Yao, A. Rumishisky, *Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning*, arXiv, 2024; doi:10.48550/arXiv.2303.15647
9. T. Shahzad, T. Mazhar, M.U. Tariq et al. *A comprehensive review of large languages models: issues and solutions in learning environments,* Discover Sustainability, Vol. 6, No. 27 (2025.). doi:10.1007/s43621-025-00815-8
10. I. Shumailov, Z. Shumaylov, Y. Zhao et al., *AI models collapse when trained on recursively generated data,* Nature, Vol. 631, 755-759 (2024.). doi:10.1038/s41586-024-07566-y
11. A. M. Turing, *Computing machinery and intelligence*, Mind, 1950, No. 59, 433-460.
12. R. Penrose, „Shadows of the Mind: A Search for the Missing Science of Consciousness," Oxford University Press, 2023.