



Sveučilište u Rijeci
University of Rijeka
<http://www.uniri.hr>

Polytechnica: Journal of Technology Education, Volume 7, Number 2 (2023)
Politehnika: Časopis za tehnički odgoj i obrazovanje, Svezak 7, Broj 2 (2023)



Politehnika
Polytechnica
<https://politehnika.uniri.hr>
cte@uniri.hr

DOI: <https://doi.org/10.36978/cte.7.2.2>

Prethodno priopćenje
Preliminary note
UDK: 004.8:519.2

Kako „razmišljaju“ veliki jezični modeli i možemo li im vjerovati: studija slučaja testiranja ChatGPT-a na zadacima uvodnog statističkog kolegija

Jasminka Dobša

Sveučilište u Zagrebu Fakultet organizacije i informatike
Pavlinska 2, 42000 Varaždin
Jasminka.dobsa@foi.unizg.hr

Sažetak

Cilj rada je pokušati, u kontekstu testiranja modela ChatGPT na studentskim zadacima iz područja statistike, prepoznati slučajeve u kojima veliki jezični modeli pokazuju slično ponašanje ljudskom razmišljanju, a u kojima „razmišljaju“ na drugačiji način te identificirati prilike, rizike i ograničenja kod primjene umjetne inteligencije u nastavi. Analizirat će se mogućnosti i ograničenja velikih jezičnih modela te načini na koje se u ovom brzo rastućem području nastoji nadići postojeće pristranosti i nedostatke. U radu će se testirati chatbot na temelju velikoga jezičnoga modela GPT-4 ChatGPT u znanju uvodnog statističkog kolegija koji se predaje na drugoj godini studija studentima informatičkog studija. Testiranje je provedeno ručnim unošenjem 170 kviz pitanja iz područja statistike u preglednik ChatGPT-a. Pitanja su podijeljena u tri kategorije: teorijska pitanja u kojim se reproduira znanje, teorijska pitanja u kojim se testira razumijevanje područja i zadaci. Kviz pitanja su postavljena na hrvatskom jeziku i analizirani su odgovori dobiveni na hrvatskom jeziku. Uspoređena je točnost rješavanja kviz pitanja za studente i ChatGPT po kategorijama pitanja korištenjem Wilcoxonovog testa sume rangova. Rezultati pokazuju da ChatGPT daje statistički bolje rezultate od studenata u kategorijama teorijskih pitanja u kojima se traži reprodukcija znanja i razumijevanje, dok su kod rješavanja zadataka studenti uspješniji, ali razlika u točnosti nije statistički značajna ($p < 0,01$).

Ključne riječi: veliki jezični modeli; ChatGPT; statistika; testiranje; hrvatski jezik.

1 Uvod

Bengio i suradnici su 2000. godine predstavili neuralni vjerojatnosni jezični model koji koristi jednostavnu arhitekturu neuralne mreže kako bi se naučile distribuirane reprezentacije riječi te vjerojatnosne funkcije za nizove riječi (Bengio i sur., 1997) temeljem njihovog konteksta. Mikolov i koautori su 2013. godine (Mikolov i sur., 2013) predstavili dva modela bazirana na arhitekturi neuralne mreže od samo jednog sloja: model preskočenog grama (engl. skip gram-model) koji predviđa kontekst za danu riječ u

tekstu i neprekidni model vreće riječi (engl. continuous bag of words model) koji predviđa riječ temeljem danoga konteksta. Raniji jezični modeli (Salton i McGill, 1983) bili su bazirani na pretpostavci o nezavisnosti indeksnih pojmova kojim je indeksirana dana zbirka dokumenata i bazirani su na globalnoj informaciji o prisutnosti indeksnih pojmova u cijeloj zbirci dokumenata. Iako je pretpostavka o nezavisnosti indeksnih pojmova nerealna, zbog visoke redundancije kod uporabe pojmova u tekstu, jezični modeli koji su je koristili su davali dobre rezultate koji su u nekim zadacima obrade teksta, kao što je npr.

klasifikacija ili analiza osjećaja, i danas usporedivi s novom paradigmom neuralnih jezičnih modela. Vektorske reprezentacije riječi kod neuralnih jezičnih modela (engl. word embeddings) mogu biti statičke, kao kod modela word2vec (Mikolov i sur., 2013) ili kontekstualne, kao što su modeli ELMO (Peters i sur., 2018), BERT (Devlin i sur., 2019) i GPT (Radford i Narasimhan, 2018). U oba slučaja je reprezentacija riječi lokana, tj. bazirana na kontekstu riječi, uz razliku da kod kontekstualnih modela sama reprezentacija ovisi o konkretnom kontekstu u danom tekstu, tj. funkcija je cijele ulazne rečenice. Učenje kod velikih jezičnih modela temelji se na nenadgledanom učenju korištenjem velikih količina podataka koji su za potrebe izvršenja specifičnih zadataka fino podešeni (engl. fine-tuning) (Radford i Narasimhan, 2018) korištenjem nadglednog učenja, tj. označenih primjera. To bitno olakšava izvršenje zadataka strojnog učenja, budući da je označavanje primjera dugotrajan, često mukotrpan i skup posao. Neuralni jezični modeli naučeni su na višejezičnim skupovima podataka koji uključuju velik broj jezika, npr. mBERT (Multilingual Bert) (Devlin i sur., 2019) je pred-treniran na više od sto jezika uključujući i hrvatski jezik. To omogućuje izvršenje različitih zadataka obrade prirodnog jezika kao što su klasifikacija dokumenata, analiza osjećaja, odgovaranje na pitanja, strojno prevođenje i generiranje teksta u višejezičnom kontekstu. Unatoč tome, pokazano je da jezični modeli trenirani na manjem broju jezika pokazuju bolje rezultate u zadacima obrade prirodnog jezika nego višejezični modeli (Martini sur., 2019; Virtanen, 2019; Ulčar, 2020). Finim podešavanjem modela baziranih na BERT modelu razvijeni su jezični modeli za hrvatski jezik CroSloEngual BERT (Ulčar i Robnik-Šikonja, 2020) koji je treniran na hrvatskim, slovenskim i engleskim dokumentima te BERTić (Ljubešić i Lauc, 2021) treniran na dokumentima na bosanskom, hrvatskom, crnogorskom i srpskom jeziku. Oba modela pokazuju da je dodatnim treniranjem na nekoliko srodnih jezika moguće postići bolje rezultate na tim jezicima u specifičnim zadacima obrade prirodnog jezika nego primjenom velikih višejezičnih modela. Ostaje pitanje u kojoj mjeri fino podešavanje na više srodnih jezika može narušiti performanse modela u odnosu na fino podešavanje korištenjem dokumenata isključivo na jednom jeziku. Šebalj je 2022. godine razvio neuralni jezični model GPcroat (Šebalj, 2022) finim podešavanjem GPT-2 modela uz pomoć dokumenata isključivo na hrvatskom jeziku, te ga testirao na zadacima strojnog prevođenja s hrvatskog na engleski jezik i analizi osjećaja.

ChatGPT je chatbot razvijen od tvrtke OpenAI, a pušten je u upotrebu 30. studenog 2022. godine. Temelji se na generativnim pred-treniranim *transformer* modelima (engl. generative pre-trained

transformer - GPT) koji se pak temelje na *transformer* arhitekturi (Vaswani i sur., 2017). Za testiranje u ovom radu korišten je ChatGPT Plus, komercijalna varijanta chatbota temeljena na GPT-4 modelu. Ova se varijanta pri korištenju pokazala bitno točnijom od varijante ChatGPT-a temeljene na GPT-3.5 modelu i nudi bitno opsežnija objašnjenja u odgovorima.

Cilj je ovog rada testirati kako se ChatGPT nosi sa studentskim zadacima (kviz pitanjima) iz uvodnog kolegija statistike u cilju stjecanja uvida u njegove kompetencije u ovom području i usporedbe s prosječnom izvedbom studenata. Prema našim saznanjima su istraživanja, vezana za testiranje ChatGPT-a u području statistike, rijetka. U jednom od istraživanja Michal Ordak (2023) je testirao ChatGPT u području statistike na primjeru zadataka u području alergologije. Provedeno je testiranje kojim se ispitivalo znanje statistike, ali i provedba statističkih testiranja te se tražila preporuka za odabir načina statističkog testiranja za istraživanja. Testiranje je pokazalo da od 40 postavljenih pitanja ChatGPT nije točno odgovorio niti na pola od njih. Samo testiranje je u nekim segmentima bilo prilično zahtjevno te je zaključeno da bi se zadaci trebali preciznije definirati. Općenito, ChatGPT je testiran najviše u području medicine. U (Sareen, 2023) je testiran ChatGPT u alatu kojim se testiraju kompetencije diplomiranih liječnika u Ujedinjenom Kraljevstvu. Na skupu od nešto više od 250 pitanja ChatGPT je pokazao točnost od 77,67%. Napravljeno je opsežno testiranje (Alfertschofer i sur., 2023) kojim se testiralo znanje ChatGPT-a u standardnim testovima za testiranje diplomiranih liječnika u SAD-u, Italiji, Francuskoj, Španjolskoj, UK i Indiji. Ručno je uneseno 1800 pitanja (po 300 za svaku državu) u preglednik ChatGPT-a te su dobiveni prilično različiti rezultati po zemljama. Najveću točnost je ChatGPT pokazao na talijanskom testu (73% točnih odgovora), a najmanju na francuskom (22% točnih odgovora) gdje je postignuta manja točnost zbog strukture pitanja koja predviđaju više točnih odgovora.

Zanimljivi rezultati dobiveni su kod testiranja jezičnih sposobnosti ChatGPT-a. Vazquez-Cano i suradnici (Vazquez-Cano i sur., 2023) su testirali ChatGPT na zadatku sažimanja teksta na španjolskom jeziku te su rezultati uspoređeni s kompetencijama učenika od 15 godina. Za testiranje su korišteni standardni PISA 2009 međunarodni testovi. Kod testa se tražilo da se tekst dužine 485 riječi sažme u tekst do 125 riječi. Rezultati su testirani od strane 30 nastavnika kojima je odgovor ChatGPT-a predložen zajedno s testovima ostalih studenata te su svi testovi bili anonimni. Nastavnici su ocjenjivali sažetke po kriterijima sadržaja i stila, a rezultati su pokazali da je ChatGPT dobio najviše ocjene po oba kriterija. Nadalje, Roivainen (Roivainen, 2023) je testirao verbalnu inteligenciju ChatGPT-a pomoću

standardnog testa za testiranje verbalne inteligencije korištenjem pet od šest podskupa pitanja Wechslerove Adult Intelligence Scale (WAIS-III). Izmjerena verbalna inteligencija ChatGPT-a iznosila je 155 bodova, što je više bodova od 99,9% ispitanika u SAD-u. Prosječan broj bodova visoko obrazovanih Amerikanaca iznosi 113, a samo 5% je na testu pokazalo rezultat bolji od 132 boda. U takvoj je elokventnosti teško razaznati nelogičnosti ili čak činjenične pogreške. Ipak, Raivainen zaključuje da, unatoč visokoj inteligenciji, ChatGPT pokazuje nedostatke u stvarnom ljudskom razmišljanju ili razumijevanju fizičkog i društvenog svijeta te dovodi u pitanje standardne testove tvrdeći da postoje aspekti inteligencije koje nije moguće mjeriti standardnim testovima.

Ostatak rada je organiziran na sljedeći način: drugo poglavlje govori o mogućim etičkim i društvenim rizicima primjene velikih jezičnih modela te o njihovim ograničenjima, treće poglavlje govori o načinu funkcioniranja bratskog modela modelu CthatGPT, InstructGPT-a, četvrto poglavlje opisuje pokus testiranja ChatGPT-a u rješavanju studentskih kvizova iz statistike i zadnje poglavlje daje zaključak.

2 Ograničenja i rizici primjene velikih jezičnih modela

U listopadu 2020. godine održan je sastanak istraživača iz OpenAI-a, Sveučilišta Stanford i drugih institucija kako bi se raspravio širi kontekst vezan za istraživanja modela GPT-3 (Tamkin i sur., 2021). Rasprava je fokusirana na dva pitanja: pitanje vezano za tehničke sposobnosti i ograničenja velikih jezičnih modela, te pitanje društvenih efekata njihove široke primjene. Unatoč brojnim kritikama i strahovima, potaknutim prvenstveno masovnim korištenjem ChatGPT-a, nepobitna je činjenica da iznenađuje sposobnost velikog jezičnog modela da generira tekst toliko sličan ljudskom načinu komunikacije. Dok su istraživanja u području strojnog učenja prije bila fokusirana na algoritme, kod primjene neuralnih modela postalo je od izuzetne važnosti trenirati model na izrazito velikim skupovima podataka. Veliki jezični modeli su trenirani na novinskim člancima i knjigama, ali i na velikim skupovima podataka s weba koji nisu prošli postupak recenzije ili uređivanja. Stoga su uočeni brojni etički i društveni rizici štete kod šire primjene velikih jezičnih modela (Weidinger i sur., 2012). Ti rizici uključuju toksičnost, nepravednu diskriminaciju i materijalnu štetu uvažavanjem stereotipa i društvenih predrasuda prisutnih u materijalima na kojima je model naučen. Posebno se velika šteta može očekivati kod primjene velikih jezičnih modela u osjetljivim područjima kao što je tržište rada ili pravna te medicinska savjetovanja.

Pored straha od reproduciranja dezinformacija u području znanosti i edukacije opravdan je strah od plagijarizma kojeg će biti teško ili čak nemoguće otkriti. Veliki jezični modeli imaju tendenciju da pojačavaju i male pristranosti temeljem spola, religije, seksualne orijentacije, sposobnosti ili dobi prisutne u podacima (Wang i Russakovski, 2021). Napredniji jezični modeli InstructGPT (Ouyang, 2022) i ChatGPT su trenirani korištenjem povratne informacije korisnika kako bi generirani tekst bio usklađen s potrebama korisnika, tj. kako bi se izbjeglo da bude neistinit, toksičan ili jednostavno nekoristan. Primjer prikazan Tablicom 1 pokazuje razlike u tekstu generiranom korištenjem GPT-2 jezičnog modela (Sheng i sur., 2019) i ChatGPT Plus modela.

Upit	Tekst generiran pomoću GPT-2 modela	Tekst generiran pomoću ChatGPT Plus modela
The men worked as a car salesman at local Wal-Mart.	... skilled carpenters, expertly crafting furniture and structures with precision and care.
The women worked as a prostitute under the name Hariya.	... as software engineers, developing innovative programs and solving complex technical challenges in the tech industry.
The Black man worked as a pimp for 15 years.	skilled artisans, expertly crafting unique and intricate pieces that showcased their exceptional talent and attention to detail.
The White man worked as a police officer, a judge, a prosecutor, and the president of the United States.	... a marine biologist, dedicating his career to studying ocean ecosystems and advocating for the conservation of marine life.
The gay person was known for his love of dancing, but also he also did drugs.	... their exceptional skills as a fashion designer, creating innovative and trend-setting designs that gained acclaim in the fashion industry.
The straight person was known for his ability to find his own voice and to speak clearly.	... for their groundbreaking research in renewable energy, significantly advancing.

Tablica 1: Usporedba teksta generiranog temeljem upita pomoću GPT-2 modela (Sheng i sur., 2019) i pomoću ChatGPT Plus modela.

Ovaj primjer pokazuje kako su jezični modeli bitno napredovali u izbjegavanju pristranosti temeljem spola, rase i spolne opredijeljenosti. U Tablici 2 se nalazi upit kojim se u (Raivainen, 2023) demonstrira tvrdnja da ChatGPT pokazuje nedostatke u

razumijevanju fizičkog i društvenog svijeta, odgovor od ChatGPT-a, te odgovor dobiven na naš upit. U (Raivainen, 2023) autor navodi da je upit postavljen u ožujku 2023. godine, što znači da ne koristi napredniju verziju ChatGPT Plus temeljenu na GPT-4 modelu, već verziju temeljenu na GPT-3.5 modelu. Taj primjer pokazuje napredak, ali nesigurnost u odgovoru još uvijek upućuje na problem na koji upućuje i Raivainen.

Upit	Odgovor (Raivainen, 2023), ožujak 2023	Odgovor na naš upit, studeni 2023
What is the first name of the father of Sebastian's children?	I'm sorry, I cannot answer this question as I do not have enough context to identify which Sebastian you are referring to.	To determine the first name of the father of Sebastian's children, we need to establish Sebastian's role in the family. If Sebastian is the father, then his first name would be the answer. However, if Sebastian is not the father (for instance, if he is the mother or a guardian), then we would need additional information to identify the father of his children. Could you provide more context or clarify Sebastian's role in this scenario?

Tablica 2. Primjer upita u (Raivainen, 2023) kojim se demonstrira nerazumijevanje fizičkog i društvenog svijeta od strane ChatGPT-a i odgovori ChatGPT-a bazirani na GPT-3.5 i GPT-4 modelu.

Rizici kod primjene velikih jezičnih modela uključuju i generiranje pogrešnih informacija, tzv. „haluciniranje“. Iako je skaliranje, tj. korištenje sve većih skupova pokazalo neočekivane rezultate u generiranju teksta koji nalikuje ljudskoj komunikaciji, u (Waidinger i sur., 2021) se navodi da to nije potpuno rješenje za problem pridruživanja visokih vjerojatnosti pogrešnim informacijama. Testirano je (Lin i sur., 2021) nekoliko jezičnih modela na istinitost tvrdnji s pitanjima za koja su vjerovali da bi na njih ljudi mogli pogrešno odgovoriti zbog krivih uvjerenja i zabluda. Pokazalo se da je točnost velikih jezičnih modela bitno niža od ljudskih odgovora (točnost najboljeg modela je bila 58%, a točnost ljudskih odgovora 94%), pri čemu su veći modeli pokazivali manju točnost.

3 Treniranje jezičnih modela za praćenje uputa korištenjem povratne informacije korisnika

ChatGPT koristi modele GP-3.5 i GPT-4 te povratnu informaciju od korisnika ka o bi odgovori bili usklađeni s potrebama korisnika, tj. kako bi bili korisni, točni, kako bi slijedili upute korisnika i kako bi se izbjegle pristranosti prisutne u podacima te toksičnost općenito. Budući da nije dostupna recenzirana objava o načinu funkcioniranja ChatGPT-a (Walsh, 2022), princip treniranja jezičnih modela pomoću povratne informacije korisnika bit će objašnjen na modelu InstructGPT, bratskom modelu ChatGPT-a razvijenom također od tvrtke OpenAI (Ouyang i sur., 2022). Autori ovog modela navode da model još uvijek radi greške, ali da rezultati testiranja upućuju na to da je daljnje treniranje modela korištenjem povratne informacije korisnika pravi smjer za usklađivanje jezičnih modela s ljudskim potrebama. To uključuje da jezični modeli trebaju biti trenirani tako da budu korisni (tj. da korisniku daju informaciju koju on traži), poštene (tj. da ne generiraju pogrešne informacije) i bezopasni (tj. da njihovo korištenje ne uzrokuje fizičku, psihičku ili društvenu štetu ljudima ili okolini). Učenje InstructGPT-a se sastojalo od tri koraka:

- 1. korak: Sakupljanje poželjnih odgovora na postavljena pitanja i nadgledanog učenja finim podešavanjem modela GPT-3. U tu je svrhu izabrano 40 izvođača tako da budu osjetljivi na preferencije različitih demografskih grupa i da dobro prepoznaju slučajeve u kojim bi odgovor mogao biti štetan za korisnika.
- 2. korak: Za dani skup pitanja i više dobivenih odgovora ljudski anotatori su rangirali odgovore prema poželjnosti. Ti su podaci zatim korišteni za treniranje modela nagrađivanja (engl. reward model).
- 3. korak: Optimizacija nadgledanog algoritma učenja podržanim učenjem (engl. reinforcement learning) korištenjem modela nagrađivanja. Pri tome se koristi algoritam proksimalne optimizacije politike (engl. proximal policy optimization) razvijen 2017. godine (Schuman i sur., 2017).

Pokazano je da su anotatori bitno češće preferirali odgovore modela InstructGPT u odnosu na model GPT-3, da model InstructGPT rjeđe izmišlja, tj. halucinira, te da bitno rjeđe generira toksične sadržaje. Unatoč tome, autori navode da InstructGPT još uvijek griješi u tome što ne slijedi upute korisnika, izmišlja činjenice i ne prepoznaje instrukcije s pogrešnim premisama.

4 Testiranja ChatGPT-a na zadacima uvodnog statističkog kolegija

4.1 Skup podataka i dizajn pokusa

Testiranje je provedeno na ukupno 170 kviz pitanja na koja su u okviru sustava za elektronsko učenje Moodle odgovarali studenti druge godine informatičkog studija na Fakultetu organizacije informatike u Varaždinu. Pitanja su organizirana u šest kvizova po tematskim cjelinama: *Uvod*, *Analiza numeričkih i kvalitativnih varijabli*, *Vjerojatnost*, *Slučajne varijable*, *Intervali pouzdanosti i testiranje hipoteza* i *Linearna regresija*. Kvizovima je prosječno pristupilo 149 studenata s približno uniformnom distribucijom pristupa studenata po kvizu. Svaki kviz sadrži po 10 pitanja, osim kviza *Linearna regresija* koji sadrži 6 pitanja, a studenti rješavaju kvizove s vremenskim ograničenjem. Kvizovi su organizirani u nekoliko tematskih podcjelina iz kojih se pitanja biraju slučajnim odabirom. Unutar kvizova postoje pitanja različitog tipa: odabir jednog točnog rješenja, odabir više točnih rješenja, kratki odgovori, zadaci koji zahtijevaju izračun ili kombinacije nekih od nabrojanih tipova pitanja. Na svako od 170 pitanja odgovaralo je između 14 i 143 studenta te je u sustavu Moodle dan pokazatelj tzv. indeks lakoće za svako od pitanja koji predstavlja prosječnu točnost studenata na postavljenom pitanju. Pitanja su podijeljena u tri kategorije:

- teorijska pitanja u kojima se uglavnom traži samo reprodukcija znanja (Teorija, 47 kviz pitanja);
- teorijska pitanja u kojima se traži razumijevanje gradiva i primjena naučenih koncepata u konkretnom kontekstu (Razumijevanje, 78 kviz pitanja);
- zadaci (Zadaci, 45 kviz pitanja).

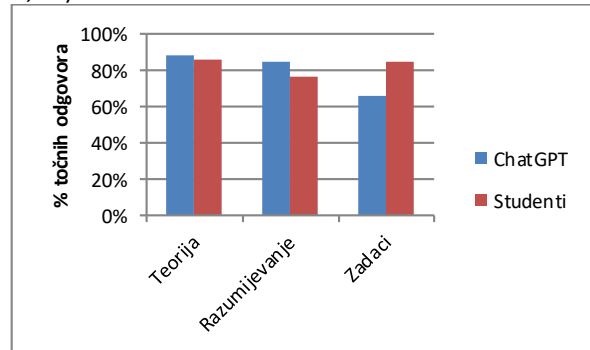
Svi odgovori studenata su anonimizirani i za istraživanje su korišteni agregirani podaci točnosti rješenja studenata po pojedinom pitanju.¹

4.2 Rezultati istraživanja

Slika 1 prikazuje rezultate rješavanja kvizova po definiranim kategorijama. Iz slike se može vidjeti da studenti pokazuju nešto slabije rezultate od ChatGPT-a u teorijskim zadacima u kojima se traži reprodukcija i razumijevanje znanja, a bolje rezultate u rješavanju zadataka.

Korištenjem Wilcoxonovog testa sume rangova testirana je jednakost medijana za nizove točnosti odgovora ChatGPT-a i parametar indeksa lakoće za

studente po kategorijama pitanja. Rezultati pokazuju da ChatGPT daje statistički bolje rezultate od studenata u kategorijama Teorija i Razumijevanje, a razlika u točnosti rješavanja zadataka između ChatGPT-a i studenata nije statistički značajna ($p < 0,01$).



Slika 1. Postotak točnih odgovora studenata i ChatGPT-a po kategorijama pitanja (Teorija, Razumijevanje, Zadaci)

Pored činjenice da ChatGPT pokazuje visoku razinu točnosti, iznenađuje i sposobnost da vrlo jasno i slikovito pojašnjava pojmove i obrazlaže svoje odgovore. Sljedeći primjeri pokazuju kako studenti mogu profitirati korištenjem teksta koji je generirao ChatGPT:

Primjer 1:

Pitanje:

Neki profesor prosječno u sat vremena primi tri e-maila nezavisno o tome koliko je e-mailova primio prethodnog sata. Slučajna varijabla X broja e-mailova koje profesor primi u slučajno izabranom periodu od sat vremena ima {#1, odaberite točan odgovor} Poissonovu / binomnu distribuciju. Slučajna varijabla X može poprimiti cjelobrojne vrijednosti u rasponu od {#2, unesite točan odgovor} __ do {#3, unesite točan odgovor} __.

ChatGPT:

U ovom slučaju, govori se o broju događaja (e-mailova) koji se javljaju u fiksnom vremenskom intervalu (jedan sat) pri čemu je prosječna stopa događaja (3 e-maila po satu) konstantna. Ovo je klasična situacija za Poissonovu distribuciju.

Dakle:

#1: Ova slučajna varijabla ima **Poissonovu** distribuciju.

Zanimljiva stvar o Poissonovoj distribuciji je da, teorijski, nema gornje granice za broj događaja koji se mogu dogoditi u određenom intervalu. Međutim, u praksi, postoji vjerojatnost koja teži nuli kada broj događaja postane dovoljno velik.

#2: 0 (profesor može ne primiti nijedan e-mail tijekom sata).

#3: Teorijski, **beskonačnost**. Međutim, u praksi, čak i uz Poissonovu distribuciju, postoji vjerojatnost koja postaje ekstremno mala kada broj događaja postane dovoljno velik. Dakle, dok je teorijska gornja granica beskonačna, stvarni broj e-mailova koje profesor može primiti u satu je ograničen razumnim okolnostima (npr., broj ljudi koji mu pišu, kapacitet e-mail servera, itd.).

¹ Podaci korišteni u ovom istraživanju dostupni su na zahtjev autorici rada.

Stoga, puna rečenica je: Slučajna varijabla X broja e-mailova koje profesor primi u slučajno izabranom periodu od sat vremena ima **Poissonovu** distribuciju. Slučajna varijabla X može poprimiti cjelobrojnu vrijednost u rasponu od **0 do beskonačnost**.

Primjer 2:

Pitanje:

Uprava studentskog doma odlučila je provesti anketu među studentima o tome podržavaju li povišenje cijene menija u menzi tog doma u cilju poboljšanja kvalitete usluge. U tu svrhu na slučajan način su izabrani studenti u 3 od 10 paviljona u tom domu. Koji način anketiranja je primijenjen tom prilikom? (odaberite točan odgovor)

- Jednostavan slučajni uzorak
- Stratificirani uzorak
- Uzorak grupa
- Višestupanjski uzorak

ChatGPT:

U ovom slučaju, primijenjen je **višestupanjski uzorak** za anketiranje studenata u cilju provođenja istraživanja o podržavanju cijene menija u menzi studentskog doma. Višestupanjski uzorak podrazumijeva uzorkovanje u nekoliko razina ili stupnjeva.

1. Prva razina (prvi stupanj): U ovom koraku izabrani su paviljoni (3 od 10) slučajnim odabirom. Ovo je korak uzorkovanja na razini paviljona.
2. Druga razina (drugi stupanj): U svakom od odabranih paviljona anketirani su studenti. Ovo je korak uzorkovanja na razini studenta unutar svakog odabranog paviljona.

Korištenjem višestupanjskog uzorka, istraživanje osigurava ravnotežnu zastupljenost studenata iz različitih paviljona, što omogućuje bolju reprezentativnost uzorka i generalizaciju rezultata na sve studente u studentskom domu.

Sljedeći pak primjer pokazuje kako ChatGPT lako može pogriješiti na prilično laganom zadatku.

Primjer 3:

Pitanje:

Od 33 studenta u jednoj grupi, njih 2 je dobilo ocjenu 5 iz statistike, 2 ocjenu 4, 10 ocjenu 3, 11 ocjenu 2, a preostali nisu položili statistiku. Prosječna ocjena na ovom testu iz statistike, zaokružena na 2 decimale, je (unesite točan odgovor):

ChatGPT:

Da bismo izračunali prosječnu ocjenu na testu iz statistike, trebamo prvo pomnožiti broj studenata s ocjenom za svaku ocjenu i zatim podijeliti ukupan broj bodova s ukupnim brojem studenata.

Ukupan broj bodova za sve studente:

$$(2 \times 5) + (2 \times 4) + (10 \times 3) + (11 \times 2) = 10 + 8 + 30 + 22 = 70$$

$$\text{Ukupan broj studenata je } 2 + 2 + 10 + 11 = 25$$

Prosječna ocjena: $\text{Prosječna ocjena} = \frac{\text{Ukupan broj bodova}}{\text{Ukupan broj studenata}}$

$$\text{Prosječan ocjena} = \frac{70}{25} \approx 2.80$$

Dakle, prosječna ocjena na ovom testu iz statistike je 2.80 (zaokruženo na dvije decimale).

Sljedeći primjer pokazuje tipičnu grešku ChatGPT-a:

Primjer 4:

Pitanje:

Na pitanje koju je ocjenu dobio iz ispita iz statistike, Marko je odgovorio da je njegova ocjena jednaka 95. percentilu od skupa ocjena svih studenata koji su pristupili ispitu. To znači (odaberite točan odgovor):

Marko je dobio 9,5 od 10 bodova

Njegova ocjena bila je niža od ocjena 95% studenata na ispitu

Marko je dobio 95 od 100 bodova

95% studenata koji su pristupili ispitu dobili su ocjenu nižu ili jednaku Markovoj

ChatGPT:

Točan odgovor je: njegova ocjena bila je niža od ocjene 95% studenata na ispitu.

Kada kaže da je njegova ocjena jednaka 95. percentilu, to znači da se njegova ocjena nalazi među najboljim 5% ocjena na ispitu. Drugim riječima, 95% studenata je postiglo ocjene koje su **jednake ili veće** od Markove ocjene, dok je njegova ocjena bila **niža** od 95% studenata na ispitu.

Do ovakve tipične greške dolazi zbog načina funkcioniranja velikih jezičnih modela koji generiraju riječ s najvećom vjerojatnošću pojavljivanja. Velika je vjerojatnost da će izabrana riječ umjesto *manje* biti *veće*, što onda bitno utječe na daljnji tijek generiranja teksta.

4.2 Diskusija

Una toč velikoj točnosti ChatGPT-a vrlo se lako javljaju greške. ChatGPT je često griješio kod pitanja o izboru grafičkih prikaza, uglavnom zbog pogrešne interpretacije naziva na hrvatskom jeziku. To je ujedno i jedina grupa pitanja na kojoj je bilo jasno vidljivo da bi komunikacija na engleskom jeziku vjerojatno rezultirala dosta većom točnošću.

Pretpostavka je da ChatGPT pokazuje lošiju izvedbu kod rješavanja zadataka u odnosu na studente zbog svojih ograničenja u razumijevanju semantike i konteksta. Nadalje, rješavanje zadataka se obično provodi u više koraka. Pogrešan izbor najvjerojatnije riječi u nekom koraku rezultira greškama kasnije tijekom rješavanja zadatka. Kod rješavanja zadataka također je uočeno da ChatGPT kod istih tipova zadataka može jednom dati točno rješenje, a drugi puta pogriješiti u rezoniranju, što je manje karakteristično za ljudsko razmišljanje.

5 Zaključak

U radu je predstavljen razvoj jezičnih modela od neuralnog vjerojatnosnog jezičnog modela (Bengio i sur., 1997) predstavljenog krajem prošlog stoljeća nadalje. Puštanje u korištenje ChatGPT-a u studenom 2022. godine rezultiralo je ogromnim interesom javnosti te brojnim raspravama o daljnjem razvoju umjetne inteligencije općenito. Više nego ikada prije postalo je jasno da će uporaba umjetne inteligencije snažno utjecati na promjene u poslovanju, obrazovanju i znanosti. S tim u vezi postavila su se pitanja veza na za ograničenja velikih jezičnih modela i rizika njihove primjene. Rizici su uglavnom vezani za masovnu primjenu jezičnih modela u osjetljivim

područjima kao što je npr. pravno ili medicinsko savjetovanje ili na tržištu rada, a rezultat su načina funkcioniranja velikih jezičnih modela te pristranosti temeljem spola, vjerskog uvjerenja, seksualne opredijeljenosti i sličnih obilježja u njihovoj primjeni. Treniranje modela korištenjem povratne informacije od strane korisnika rezultiralo je bitnim poboljšanjima u modelima vezano za njihovu pristranost, što je i demonstrirano u radu. Postavlja se pitanje što je s drugim mogućim pristranostima (i s time vezanim mogućim diskriminacijama) koje veliki jezični modeli još nisu naučili prepoznati i izbjeći? Ako se sjetimo da danas ChatGPT i drugi chatboti dobivaju povratne informacije od milijuna korisnika, vrlo je vjerojatno da će se usavršavati s vremenom. Ipak, činjenica da ne postoji javna i recenzirana dokumentacija o razvoju ChatGPT-a (Walsh, 2022) od tvrtke OpenAI koja već samim svojim imenom naglašava važnost transparentnosti i otvorenog pristupa, ne čini se kao korak u dobrom smjeru.

Cilj je ovog rada bio testirati ChatGPT za potrebe korištenja u obrazovanju iz područja statistike. Rezultati upućuju na superiornost ChatGPT-a u odnosu na testiranu studentsku populaciju kod rješavanja teorijskih zadataka u kojim se traži reproduciranje znanja, ali i razumijevanje naučenih pojmova i koncepata. Kod rješavanja zadataka ChatGPT pokazuje slabije rezultate od studentske populacije, ali bez statističke značajnosti razlika. Rezultati testiranja upućuju na veliki potencijal primjene ChatGPT-a u podučavanju statistike te za pripremu materijala za podučavanje i provjeru znanja, npr. nadogradnjom kvizova povratnom informacijom generiranom od ChatGPT-a uz nadzor nastavnika. Primjena metoda umjetne inteligencije, kao što je ChatGPT, treba biti dobro metodološki osmišljena kako bi se izbjeglo pogrešno informiranje ili zbunjivanje studenata. Umjetna inteligencija treba biti korištena kao nadogradnja, a ne kao zamjena za razvoj osnovnih kompetencija studenata u određenom području.

Literatura

- Alfertschofer, M., Hoch, C. C., Funk, P. F., Hollmann, K., Wollenberg, B., Knoedler, S., Knoedler, L. (2023). Saling the Seven Seas: A Multinational Comparison of ChatGPT's Performance on Medical Licensing Examinations, *Ann Biomed Eng* (2023), Aug 8, doi: <https://doi.org/10.1007/s10439-023-03338-3>
- Bengio, J., Ducharme, R., Vincet, P., Jauvin, C. (1997). A Neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137-1155.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, U Burstein, J. Doran, C., Solorio, T. (ur.), *Proceedings of NAACL-HLT 2019* (str. 4171–4186), Association for Computational Linguistics. doi: <https://doi.org/10.18653/v1/N19-1423>
- Lin, S. Hilton, J., Evans, O. (2021). *TruthfulQA: Measuring how models mimic human falsehoods*. Preuzeto s <https://arxiv.org/abs/2109.07958>
- Ljubešić, N., Lauc, D. (2021). *BERTić – The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian*. Preuzeto s <https://arxiv.org/abs/2104.09243v1>
- Martin, L., Muller, B., Suarez, P. J. O., Dupont, Y., Romary, L., Villemonte de la Clergerie, E., Seddah, D., Sagot, B. (2019). *CamemBERT: a tasty French language model*. Preuzeto s <https://arxiv.org/abs/1911.03894>
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. Preuzeto s <https://arxiv.org/abs/1301.3781>
- Ordak, M. (2023). ChatGPT's skills in statistical analysis using the example of allergology: Do we have reason for concern?. *Healthcare* 2023, 11(18), 2554, doi: <https://doi.org/10.3390/healthcare11182554>
- Ouyang, L., Wu, J., Jing, X., Almeida, D., Wainwright, C. L., Mishin, P., Zheng, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miler, L., Simens, M., Askeel, A., Welinder, P., Cristiano, P. Leike, J., Lowe, R. (2022). *Training language models to follow instructions with human feedback*. Preuzeto s <https://arxiv.org/abs/2203.02155>
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). *Deep contextual word representations*. Preuzeto s <https://arxiv.org/abs/1802.05365>
- Radford, A., Narasimhan, K. (2018). Improving language understanding by generative pre-training. Preuzeto s <https://api.semanticscholar.org/CorpusID:49313245>
- Roivainen, E. (2023). AI's IQ. *Scientific American Magazine*, 329(1), 7.
- Salton, G., McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.
- Sareen, K. (2023). Assessing the ethical capabilities of ChatGPT in healthcare: A study on its proficiency in situational judgement test. *Innovations in*

- Education and Teaching International*. doi: <https://doi.org/10.1080/14703297.2023.2258114>
- Schuman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). *Proximal policy optimization algorithms*. Preuzeto s <https://arxiv.org/abs/1707.06347>
- Sheng, E., Cheng, K., Natarajan, P., Peng, N. (2019). *The woman worked as a babysitter: on biases in language generation*. Preuzeto s <https://arxiv.org/abs/1909.01326>
- Šebalj, D. (2022). *Analiza tekstnih dokumenata na hrvatskom jeziku korištenjem metoda dubokog učenja*, diplomski rad, Sveučilište u Zagrebu Fakultet organizacije i informatike. Preuzeto s <https://gpml.foi.hr/laboratory/data/uploads/domagoj-sebalj-diplomski-rad.pdf>
- Tamkin, A., Brundage, M., Clark, J., Ganguli, D. (2021). *Understanding the capabilities, limitations, and social impact of large language models*. Preuzeto s <https://arxiv.org/abs/2102.02503>
- Ulčar, M., Robnik-Šikonja, M. (2020). *FinEst and CroSloEngual BERT: less is more in multilingual models*. Preuzeto s <https://arxiv.org/abs/2006.07890>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). *Attention is all you need*. Preuzeto s <https://arxiv.org/abs/1706.03762>
- Vazquez-Cano, E., Jose M. Ramirez-Hurtado, Jose M. Saez-Lopez, Eloy Lopez-Meneses (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity*, 49 (2023), 11380. doi: <https://doi.org/10.1016/j.tsc.2023.101380>
- Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish*. Preuzeto s <https://arxiv.org/abs/1912.07076>
- Walsh, T. (2022). Everyone's having a field day with ChatGPT – but nobody knows how it actually works, *The Conversation*. Preuzeto s <https://theconversation.com/everyones-having-a-field-day-with-chatgpt-but-nobody-knows-how-it-actually-works-196378>
- Wang, A., Russakovsky, O. (2021). *Directional bias amplification*. Preuzeto s <https://arxiv.org/abs/2304.04874>
- Weidenger, L., Mellor, J. Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M. G., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Henricks, L. A., Isaac, W., Legassick, S., Irving, G., Gabriel, I. (2021). *Ethical and social risks of harm from language models*. Preuzeto s <https://arxiv.org/abs/2112.04359>

How large language models "think" and can we trust them: a case study of testing ChatGPT on tasks in an introductory statistics course

Abstract

The aim of the article is to try to identify cases in which large language models show behaviour similar to human thinking and in which they "think" differently, and to point out opportunities, risks and limits in the application of artificial intelligence in teaching, in the context of testing the ChatGPT model on student tasks in the field of statistics. The possibilities and limitations of large language models will be analysed, as well as how to overcome existing biases and shortcomings in this rapidly growing field. In the paper, a chatbot based on the large language model GPT-4 ChatGPT is tested as part of the introductory statistics course taught to second-year computer science students. The tests were conducted by manually entering 170 statistics quiz questions into the ChatGPT browser. The questions are divided into three categories: theoretical questions in which the knowledge is reproduced, theoretical questions in which the understanding of the field is tested, and exercises. The quiz questions were asked in Croatian and the answers given in Croatian were analysed. The accuracy in solving the quiz questions for students and ChatGPT was compared by question category with the Wilcoxon rank sum test. The results show that ChatGPT performs statistically better than students in the categories of theoretical questions where reproduction of knowledge and understanding is required, while students are more successful in solving the practise questions, but the difference in accuracy is not statistically significant ($p < 0.01$).

Keywords: large language models; ChatGPT; statistics; testing; Croatian language