Korištenje velikih jezičnih modela u ocjenjivanju zadataka otvorenog tipa

Runjić, Lucija

Undergraduate thesis / Završni rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Electrical Engineering and Computing / Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:168:562092

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-23**

Repository / Repozitorij:

FER Repository - University of Zagreb Faculty of Electrical Engineering and Computing repozitory

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1280

KORIŠTENJE VELIKIH JEZIČNIH MODELA U OCJENJIVANJU ZADATAKA OTVORENOG TIPA

Lucija Runjić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1280

KORIŠTENJE VELIKIH JEZIČNIH MODELA U OCJENJIVANJU ZADATAKA OTVORENOG TIPA

Lucija Runjić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1280

Pristupnica:	**Lucija Runjić (0036538843)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentorica:	prof. dr. sc. Ljiljana Brkić

Zadatak: **Korištenje velikih jezičnih modela u ocjenjivanju zadataka otvorenog tipa**

Opis zadatka:

Problemi u stvarnom svijetu nisu uvijek dobro definirani ili ograničeni, a čak i kada jesu, često imaju više mogućih, točnih rješenja. Često ne postoje uspostavljeni formalni modeli za pružanje rješenja, a ti problemi često zahtijevaju nestrukturiran odgovor studenta, koji uključuje eseje, slike, video zapise, dijagrame i sličan sadržaj. Takvi se problemi ili zadaci nazivaju zadacima otvorenog tipa. Za razliku od zadataka zatvorenog tipa koji obično imaju jedan točan odgovor, ocjenjivanje zadataka otvorenog tipa pokazalo se teškim za automatizirati. Pouzdanim i učinkovitim u ocjenjivanju i pružanju formativne i sumativne povratne informacije za zadatke otvorenog tipa pokazalo se vršnjačko ocjenjivanje (eng. Peer Assessment). U ovom završnom radu potrebno je ispitati mogućnosti velikih jezičnih modela (eng. Large Language Model - LLM) u ocjenjivanju zadataka otvorenog tipa. Konkretno, zadacima u kojima temeljem opisa nekog poslovnog procesa treba oblikovati ER i relacijski model podataka te rješenje dostaviti u obliku dijagrama i programskog koda. Potrebno je provesti ocjenjivanje i prikupiti povratne informacije za reprezentativan broj studentskih rješenja u obliku dijagrama i programskog koda (rješenja zadataka s laboratorijske vježbe iz uvodnog predmeta iz baza podataka) koristeći LLM. Nakon toga, potrebno je analizirati kvalitetu i relevantnost ocjena i povratnih informacija koje pruža LLM. Temeljem proučavanja postojeće literature predložite i implementirajte način na koji ćete procijeni kvalitetu i relevantnost ocjena i povratnih informacija dobivenih od LLM-a. Potom usporedite ocjene dobivene od LLM-a s ocjenama koje su dodijelili drugi studenti u procesu vršnjačkog ocjenjivanja te s ocjenama koje su dodijeliti nastavnici te povratne informacije dobivene od LLM-a s povratnim informacijama koje su dodijelili studenti u vršnjačkom ocjenjivanju. Predložite mjeru ili više njih kojom bi se ocijenila sposobnost ocjenjivanja zadataka ovog tipa od strane LLM-ova.

Rok za predaju rada: 14. lipnja 2024.

Zahvaljujem se mentorici prof. dr. sc. Ljiljani Brkić na pomoći, strpljenju i korisnim savjetima pri izradi ovog rada.

Sadržaj

Uvod **1**

1. Veliki jezični modeli **2**

1.1. ChatGPT 3

1.2. Bard/Gemini 4

2. Statistički pokazatelji analize podataka **6**

2.1. Suma kvadrata 6

2.1.1. Ukupna suma kvadrata 6

2.1.2. Suma kvadrata regresije 7

2.1.3. Suma kvadrata pogreške 7

2.1.4. Veza između suma kvadrata 8

2.2. Koeficijent determinacije 8

3. Automatizirana evaluacija studentskih rješenja **10**

4. Korištene tehnologije i alati **12**

5. Dostupni materijali i metodologija **13**

5.1. Opis podatkovnog skupa 13

5.2. Metodologija 15

6. Interpretacija rezultata prikupljenih od LLM-ova **18**

6.1. Usporedba prosječnih ocjena 18

6.2. Box-plot vizualizacija 19

6.3. Numeričke vrijednosti 20

6.4. Frekvencije ocjena 21

6.5. Usporedba po pitanjima 22

7. Analiza komentara **24**

7.1. Analiza komentara za odlično riješene radove 24

7.2. Analiza komentara za radove s dosta grešaka 26

8. Evaluacija ocjenjivanja zadataka otvorenog tipa 29

9. Algoritam za dodjeljivanje bodova na temelju kalibracijskog testa 31

9.1. Dodjeljivanje bodova za ocjenjivanje 31

9.2. Dodjeljivanje bodova za zadatak otvorenog tipa 33

Zaključak 37

Literatura 38

Uvod

U posljednjih nekoliko godina, sve više napreduje razvoj velikih jezičnih modela (*engl. Large Language Models, skraćeno LLM*). Ovi modeli pokazuju izvanredne sposobnosti generiranja kvalitetnih tekstova te analize i rješavanja raznih vrsta zadataka. Upravo to ih čini sve korištenijima u područjima kao što su strojno prevođenje, obrada prirodnog jezika ili razvoj chatbotova i drugih aplikacija. Iako su sposobnosti velikih jezičnih modela impresivne, njihova primjena u ocjenjivanju zadataka otvorenog tipa donosi razne prepreke i probleme.

Zadatci otvorenog tipa imaju jako širok spektar potencijalno točnih rješenja što otežava automatiziranje njihovog ocjenjivanja. Osim toga, takvi zadatci zahtijevaju dublju analizu i poznavanje problema te kreativnost. Fokus ovog rada je istraživanje mogućnosti LLM-ova u ocjenjivanju takvih zadataka, posebno onih koji zahtijevaju oblikovanje ER i relacijskog modela podataka te dostavu rješenja u obliku dijagrama i programskog koda. Ovakvi zadatci zahtijevaju posebne analitičke sposobnosti LLM-ova kako bi mogli prepoznati strukture podataka i veze među njima te ocijeniti njihovu ispravnost.

Korišten je reprezentativan uzorak studentskih rješenja za oblikovanje modela podataka. Prikupljene su ocjene i komentari LLM-ova za studentske radove te je napravljena usporedba s ocjenama i komentarima dodijeljenim od strane drugih studenata u procesu vršnjačkog ocjenjivanja, kao i s ocjenama nastavnika. U cilju je bilo detaljno analizirati kvalitetu i relevantnost ocjena te predložiti mjere za procjenu sposobnosti ocjenjivanja LLM-ova u zadacima ovog tipa.

U ovom radu korišteni su veliki jezični modeli ChatGPT **[1]** i Bard/Gemini **[2]**. Istraživanje mogućnosti primjene velikih jezičnih modela u ocjenjivanju zadataka otvorenog tipa te analiza kvalitete i relevantnosti ocjena i komentara dobivenih putem LLM-ova pruža dublji uvid u njihovu primjenjivost i ograničenja u analizi složenih struktura podataka, kao što su ER i relacijski modeli podataka. Istraživanje ovog područja može rezultirati unaprjeđenjem procesa ocjenjivanja te tako poboljšanjem kvalitete obrazovanja i evaluacije učenja.

1. Veliki jezični modeli

Veliki jezični modeli predstavljaju vrhunski alat u području umjetne inteligencije, s posebnim naglaskom na generiranje prirodnog jezika i obavljanje raznih zadataka vezanih uz procesiranje prirodnog jezika (NLP). Njihove sposobnosti obuhvaćaju prevođenje između različitih jezika, analizu teksta, vođenje tekstualnih konverzacija. Arhitektura LLM-ova temelji se na neuronskim mrežama koje koriste metode dubokog učenja, a trenirani su na ogromnim skupovima podataka. Količina parametara takvih modela može varirati od nekoliko milijuna do nekoliko milijardi, što im omogućava generiranje teksta koji je u većini slučajeva nerazlučiv od onoga kojeg su napisali ljudi.

LLM-ovi su često preveliki da bi bili pokrenuti na jedinstvenom računalu, stoga se najčešće pružaju kao usluga putem različitih API-jeva ili mrežnih sučelja. Tako korisnici mogu pristupiti njihovim funkcionalnostima bez potrebe za vlastitom infrastrukturom za izvođenje tih modela.

Način na koji jezični modeli rade je vrlo kompleksan proces, no može ga se razbiti na nekoliko manjih faza. U prvoj fazi, modeli uče čitanjem ogromnih količina tekstova dostupnih putem Interneta. Ovaj proces uključuje analizu i obradu tekstualnih podataka kako bi model mogao prepoznati određene obrasce u jeziku. Zatim, rečenice se razbijaju na manje dijelove, poput riječi ili fraza, što modelu pomaže u radu s jezikom i shvaćanju različitih konteksta. Nakon generalnog učenja, modeli se mogu dodatno trenirati na specifičnijim područjima. Na primjer, mogu se specijalizirati za zadatke poput odgovaranja na pitanja, pisanja eseja na određene teme ili generiranja programskog koda. Ova specijalizacija omogućava modelima da postanu vrlo učinkoviti u obavljanju specifičnih zadataka, čime se značajno poboljšava njihova korisnost u praktičnim primjenama. **[3].**

Primjena LLM-ova je izrazito raznolika te može unaprijediti efikasnost u različitim aspektima našeg života. Na primjer, u poslovnom svijetu mogu se koristiti za analizu tržišta, automatsko generiranje izvješća i pružanje korisničke podrške. U obrazovanju, LLM-ovi mogu pomoći pri ocjenjivanju, pružanju personalizirane podrške učenicima i razvoju obrazovnih materijala. Također, u zdravstvu, LLM-ovi mogu pomoći u analizi medicinskih zapisa, pružanju podrške liječnicima prilikom postavljanja dijagnoza i generiranju sažetaka istraživačkih radova.

Osim toga, veliki jezični modeli igraju ključnu ulogu u razvoju novih tehnologija poput glasovnih asistenata ili chatbot-ova. Ovi sustavi koriste LLM-ove kako bi pružili prirodnu i učinkovitu interakciju s korisnicima, omogućavajući im da koriste tehnologiju na intuitivan način. U budućnosti, očekuje se da će LLM-ovi igrati još veću ulogu u digitalnoj transformaciji, pružajući napredne alate za automatizaciju i optimizaciju raznih procesa u različitim industrijama.

U nastavku je pružena detaljnija analiza "chatbotova" temeljenih na LLM-u korištenih u ovom radu, konkretno ChatGPT-a i Bard-a. Jedna od njihovih ključnih značajki je sposobnost razumijevanja i korištenja konteksta kroz cijeli razgovor. To znači da prilikom svakog korisničkog upita, modeli uzimaju obzir prethodne dijelove razgovora kako bi bolje razumjeli kontekst i pružili relevantnije odgovore. Ova im sposobnost omogućava da vode konzistentan i koherentan dijalog, čineći interakciju prirodnijom i korisnijom.

1.1. ChatGPT

ChatGPT je napredni chatbot razvijen od strane tvrtke OpenAI, baziran na velikom jezičnom modelu GPT. Ovaj alat omogućava korisnicima da oblikuju i usmjeravaju razgovor prema željenoj duljini, formatu, stilu pisanja, razini detalja i jeziku. ChatGPT je izuzetno fleksibilan u svojim mogućnostima, pružajući korisnicima priliku za interakciju na način koji najbolje odgovara njihovim potrebama.

ChatGPT je značajno unaprijeđen u odnosu na svog prethodnika, InstructGPT-a, posebno u smislu smanjenja štetnih i neistinitih informacija. InstructGPT je imao tendenciju doslovnog tumačenja i odgovaranja na upite bez prepoznavanja protučinjenične prirode određenih tvrdnji. Na primjer, ako bi mu se zadala naredba poput „Pohod rimskog carstva u 2020.-oj godini“, InstructGPT bi tu instrukciju interpretirao kao stvarnu i generirao odgovor u skladu s njom. Nasuprot tome, ChatGPT je dizajniran da prepozna protučinjenične ili nerealne tvrdnje te da pruži odgovore koji su u skladu sa stvarnošću.

Ovaj napredak u prepoznavanju i reagiranju na protučinjenične tvrdnje čini ChatGPT pouzdanijim i sigurnijim za korištenje u raznim aplikacijama. Također, smanjenje generiranja štetnih informacija povećava njegovu primjenjivost u obrazovnim i profesionalnim okruženjima gdje je točnost i pouzdanost informacija od ključne važnosti.

Uz sve ove značajke, ChatGPT ima potencijal za integraciju s raznim tehnološkim rješenjima. Na primjer, može se koristiti kao pomoćnik u softverskom razvoju za generiranje i optimizaciju koda, ili kao alat za kreatore sadržaja, pomažući u pisanju blogova, članaka i marketinških materijala. U području korisničke podrške, ChatGPT može značajno smanjiti opterećenje na ljudske agente pružajući prve linije pomoći i odgovarajući na često postavljana pitanja.

Dodatno, ChatGPT ima potencijal za poboljšanje pristupačnosti tehnologije. Može se koristiti za kreiranje prilagođenih sučelja koja pomažu osobama s invaliditetom u interakciji s digitalnim sadržajem.

ChatGPT također može igrati veliku ulogu u istraživanju i razvoju. Može se koristiti za analizu velikih količina podataka i isticanje korisnih informacija, čime se ubrzava proces donošenja odluka u znanstvenim i industrijskim istraživanjima.

1.2. Bard/Gemini

Bard, koji je u novije vrijeme poznat i kao Gemini, je chatbot razvijen od strane Google-a. Ovaj model temelji se na istoimenom velikom jezičnom modelu i predstavlja direktni odgovor na OpenAI-jev ChatGPT. Bard je dizajniran da pruži korisnicima slične mogućnosti kao ChatGPT, uključujući generiranje prirodnog jezika, vođenje tekstualnih konverzacija i obavljanje raznih zadataka obrade prirodnog jezika.

Unatoč visokim očekivanjima, Bard je naišao na mlak odaziv korisnika. Jedan od glavnih razloga za to je činjenica da su korisnici primijetili da model često generira povijesno netočne informacije i slike. Na primjer, Bard je povijesnim licima često krivo dodjeljivao boju kože i nacionalnost, što je izazvalo zabrinutost oko točnosti i pouzdanosti generiranih sadržaja. Ove netočnosti ukazale su na probleme u treniranju i testiranju modela.

Ovi problemi su vjerojatno rezultat činjenice da je Bard razvijen u užurbanim okolnostima i požureno pušten u uporabu. Google je, u nastojanju da što prije konkurira OpenAI-u i ne stvara zaostatak na tržištu, ubrzao razvoj i lansiranje Bard-a. Takav pristup je doveo do potencijalno nedovoljnog testiranja i provjere kvalitete, što je rezultiralo povećanim brojem grešaka i netočnosti u generiranim odgovorima.

Iako je Bard inicijalno razvijen kao tekstualni chatbot, Google nastavlja istraživati mogućnost integracije s drugim medijima kao što su glasovi i slike kako bi proširio svoje mogućnosti i poboljšao korisničko iskustvo.

Još jedna od ključnih značajki Bard-a jest njegova sposobnost kontinuiranog učenja. Model je dizajniran tako da analizira korisničke interakcije i povratne informacije kako bi poboljšao svoje odgovore i prilagodio se korisničkim potrebama tijekom vremena. Kroz analizu razgovora koji je u tijeku, Bard nastoji filtrirati informacije koje smatra da su korisne za korisnika u tom kontekstu.

Uz to, Bard se nastoji rasprostraniti na što više platformi i integrirati se u što više različitih aplikacija kako bi omogućili korisnicima pristup iz različitih okruženja i uređaja prema vlastitim željama. Također, Bard nudi potpunu podršku drugim aplikacijama i platformama za pristup putem API-ja, što omogućava razvojnim timovima i tvrtkama integraciju Bard-a u svoje proizvode i usluge kako bi poboljšali korisničko iskustvo i dodali vrijednost svojim proizvodima.

Unatoč početnim poteškoćama, Google nastavlja raditi na unapređenju Bard-a kako bi poboljšao njegovu točnost i pouzdanost što će doprinijeti širem prihvaćanju i korištenju ovog chatbota u različitim sferama života.

2. Statistički pokazatelji analize podataka

U nastavku je opisano nekoliko ključnih statističkih pokazatelja koji se kasnije koriste u analizi podataka te implementaciji algoritama.

2.1. Suma kvadrata

U statistici je „suma kvadrata“ jedan od korištenijih koncepata u različitim analizama. Tri osnovne komponente sume kvadrata su: ukupna suma kvadrata, suma kvadrata regresije i suma kvadrata pogreške. Svaka od ovih mjera bilježi različite aspekte varijabilnosti u podatcima.

2.1.1. Ukupna suma kvadrata

Ukupna suma kvadrata *(engl. Total Sum of Squares, skraćeno SST)* mjeri sveukupnu varijabilnost u zavisnoj varijabli unutar skupa podataka. Kvantificira ukupnu varijaciju u podatcima oko njihove srednje vrijednosti te se računa bez obzira na prisustvo ili odsustvo objašnjavajućih varijabli (nezavisnih varijabli). Dakle, SST ne uzima u obzir nikakve objašnjavajuće varijable, već računa razliku između stvarnih promatranih vrijednosti zavisne varijable i aritmetičke sredine svih vrijednosti te varijable.

SST se računa prema izrazu (1). Prvo se odredi aritmetička sredina zavisne varijable unutar skupa podataka. Zatim se, za svaku pojedinačnu promatranu vrijednost zavisne varijable, računa kvadratna razlika između te vrijednosti i prosječne vrijednosti. Konačno se sve te kvadratne razlike zbroje kako bi se dobila ukupna suma kvadrata.

$$\mathrm{SST} = \sum_{i=0}^{n} (y_i - \overline{y})^2 \tag{1}$$

2.1.2. Suma kvadrata regresije

Suma kvadrata regresije *(engl. Sum of Squares Regression, skraćeno SSR)* mjeri varijabilnost u zavisnoj varijabli koja je objašnjena modelom. SSR daje uvid koliko dobro naš model prilagođava podatke. Drugim riječima, mjeri koliko varijabilnosti u zavisnoj varijabli možemo objasniti s pomoću nezavisnih varijabli u regresiji. To je korisno za procjenu učinkovitosti nekog modela.

Pri računanju SSR-a, za svaku predviđenu vrijednost se računa kvadratna razlika između te vrijednosti i srednje aritmetičke vrijednosti zavisne varijable. Nakon toga se sve te kvadratne razlike zbroje kako bi se dobila regresijsku sumu kvadrata, kao što je prikazano u izrazu (2).

$$\mathrm{SSR} = \sum_{i=0}^{n} (\hat{y}_i - \overline{y})^2 \tag{2}$$

2.1.3. Suma kvadrata pogreške

Suma kvadrata pogreške *(engl. Sum of Squares Error, skraćeno SSE)* omogućuje kvantifikaciju varijabilnosti u zavisnoj varijabli koja nije objašnjena modelom. To je suma kvadratnih odstupanja između stvarnih i predviđenih vrijednosti. Pri bilo kakvoj analizi podataka, uvijek postoji razina neobjašnjene varijabilnosti u podacima, a SSE omogućuje mjerenje te neobjašnjene varijabilnosti. SSE pokazuje koliko su stvarne vrijednosti zavisne varijable udaljene od onih koje model predviđa.

Kod računanja, koriste se stvarne vrijednosti zavisne varijable iz skupa podataka i odgovarajuće predviđene vrijednosti koje daje model. Kao što prikazuje izraz (3), za svaku vrijednost se računa kvadratna razlika između stvarne vrijednosti i predviđene vrijednosti. Zatim se te kvadratne razlike zbrajaju kako bi se dobila suma kvadrata pogreške.

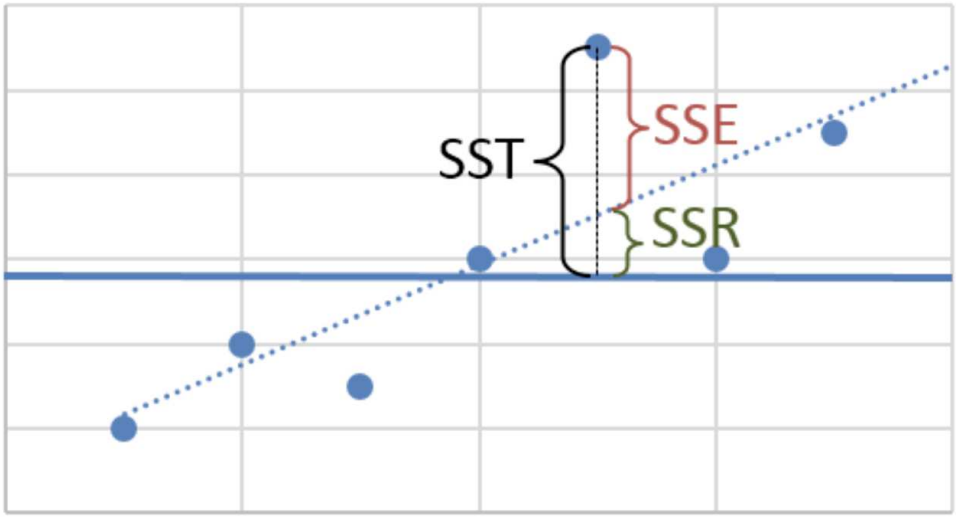
$$\mathrm{SSE} = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

2.1.4. Veza između suma kvadrata

Veza između SST, SSR i SSE kvadrata može se opisati kao u izrazu (4).

$$SST = SSR + SSE \tag{4}$$

Na *Slika 2.1* može se vidjeti grafički prikaz gore objašnjenog koncepta. Svaka točka predstavlja pojedinačnu vrijednost zavisne varijable. SSE prikazuje razliku između stvarne vrijednosti i vrijednosti predviđene modelom. SSR prikazuje razliku između vrijednosti predviđene modelom i aritmetičke sredine vrijednosti zavisne varijable. SST, koji obuhvaća SSR i SSE, prikazuje ukupnu razliku između stvarnih vrijednosti zavisne varijable i njihove aritmetičke sredine.



Slika 2.1 - Grafički prikaz veze između suma kvadrata ***[5]***

2.2. Koeficijent determinacije

Koeficijent determinacije je mjera koja se u statističkoj analizi najčešće koristi za procjenu učinkovitosti modela. Pokazuje koliki udio varijacije zavisne varijable može biti objašnjen nezavisnom varijablom modelu. Drugim riječima, predstavlja korespondenciju između zavisne i nezavisne varijable.

Računa se prema izrazu (5), a poprima vrijednosti između 0 i 1 pri čemu:

- **0** znači da model ne objašnjava nikakvu varijaciju zavisne varijable. U ovom slučaju nezavisne varijable uopće u ne pridonose predviđaju zavisne varijable što nam govori da model nije koristan
- **1** znači da model objašnjava svu varijaciju zavisne varijable što znači da sve promjene u zavisnoj varijabli mogu biti savršeno predviđene promjenama u nezavisnoj varijabli. U ovom slučaju je mjera SSR jednaka mjeri SST, a s obzirom na to da je SST zbroj SSR i SSE, to znači da je suma kvadrata pogreške jednaka nuli

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \tag{5}$$

3. Automatizirana evaluacija studentskih rješenja

Kako bi se ubrzao proces ocjenjivanja studentskih rješenja, često se koristi automatizacija uz pomoć različitih algoritama i sustava. Ova metoda ne samo da ubrzava vrednovanje, već često i smanjuje subjektivnost koja može nastati pri ručnom ocjenjivanju. Automatizirano ocjenjivanje može se primijeniti na različite vrste zadataka, no praksa je pokazala da je puno lakše automatizirati ocjenjivanje zadataka zatvorenog tipa nego zadataka otvorenog tipa.

Zadaci zatvorenog tipa su oni koji imaju unaprijed određene točne odgovore ili ograničen broj ispravnih odgovora. Primjeri ovakvih zadataka uključuju: pitanja s točno/netočno odgovorima, pitanja s višestrukim izborom odgovora te pitanja s kratkim odgovorima koji se sastoje od jedne ili najviše par riječi.

Sve je više dostupnih alata koji olakšavaju automatizaciju ovakvih zadataka. Primjerice, u zadacima s odabirom odgovora, sustav uspoređuje studentov odgovor s predefiniranim točnim rješenjem, dok kod zadataka s kratkim odgovorima sustav analizira ključne riječi ili kratke fraze.

S druge strane, zadatci otvorenog tipa nemaju unaprijed određene točne odgovore. Mogu biti vrednovani na temelju različitih kriterija. Primjeri ovakvih zadataka su: eseji, različiti projekti, „case study“ zadatci u kojima se analizira određena situacija ili problem te predlaže rješenje, evaluacija programskih rješenja kao i zadatci koji zahtijevaju primjenu znanja i vještina na specifičan problem.

Automatizacija evaluacije ovakvog tipa zadataka pokazala se kao puno veći izazov u usporedbi s automatizacijom evaluacije zatvorenog tipa. Ljudski jezik je jako kompleksan stoga je širok spektar mogućih odgovora. Za ocjenjivanje ovakvih zadataka često je potrebno razumjeti kontekst problema ili teorijsku podlogu iza rješenja, što nije jednostavno prenijeti na računalne algoritme. Vrednovanje ovakvih zadataka iziskuje subjektivnost i fleksibilnost, kako bi se rad mogao vrednovati ovisno o kontekstu zadatka te kako bi se različita rješenja prihvatila kao točna.

Ipak, postoje neke primjene ocjenjivanja zadataka otvorenog u tipa. Na primjer, postoje alati koji mogu vrednovati programska rješenja studenata tako da prvo analiziraju točnost

sintakse rješenja, a zatim testiraju učinkovitost rješenja na skupu testnih primjera. Također postoje alati koji omogućavaju grubo ocjenjivanje eseja korištenjem umjetne inteligencije. To rade analizirajući njihovu strukturu, vokabular te tražeći gramatičke pogreške.

U posljednje vrijeme, pokušava se integrirati velike jezične modele poput ChatGPT-a ili Bard-a u različite sustave za ocjenjivanje zbog njihove sposobnosti razumijevanja ljudskog jezika i generiranja odgovora na njega. Ti modeli mogu analizirati složenije odgovore i pružiti korisne sugestije za poboljšanje. **[4]**

4. Korištene tehnologije i alati

U ovom radu korištene su različite tehnologije i alati za analizu podataka, implementaciju algoritama te vizualizaciju rezultata.

Za vizualizaciju rezultata analize korišten je **R Studio IDE [6]**, integrirano razvojno okruženje (*engl. integrated development environment, skraćeno IDE*) za programski jezik **R [7]**. R Studio omogućava jednostavno pisanje i izvršavanje R skripti. Korištenje paketa **Ggplot2 [8]**, omogućava kreiranje raznih grafičkih prikaza koji olakšavaju interpretaciju podataka.

Za analizu podataka i implementaciju algoritama korišten je **PyCharm [9]**, integrirano razvojno okruženje (*engl. Integrated Development Environment, skraćeno IDE*) za programski jezik **Python [10]**. PyCharm pruža funkcionalnosti za kodiranje, debugging i testiranje Python koda. Korištenjem biblioteke P**andas [11]** i **Numpy [12]** omogućena je obrada skupova podataka te primjena statističkih metoda na iste.

Microsoft Excel [13] je korišten za pregled i početnu obradu podataka. Excel omogućava jednostavno učitavanje i manipulaciju podacima, kao i osnovnu analizu. Ovaj alat je bio korišten za pregledavanje struktura podatkovnih skupova, identifikaciju nedostajućih vrijednosti te inicijalno uređivanje podataka prije daljnje obrade u R Studiju i PyCharm-u.

5. Dostupni materijali i metodologija

5.1. Opis podatkovnog skupa

Kako bi se provela analiza ocjenjivanja zadataka otvorenog tipa od strane velikih jezičnih modela, prikupljeni su podatci o rezultatima jedne laboratorijske vježbe iz predmeta Baze podataka na drugoj godini Fakulteta elektrotehnike i računarstva.

Laboratorijska vježba sastojala se od dva dijela. U prvom dijelu su studenti dobili zadatak na temelju kojeg su trebali modelirati bazu podataka te kao rješenje predati ER model, relacijski model i SQL skriptu. U drugom dijelu laboratorijske vježbe studenti su ocjenjivali radove drugih studenata. Svaki student je dobio pet radova na ocjenjivanje, od kojih su četiri bila radovi njihovih kolega, a jedan rad je bio kalibracijski. Kalibracijski rad je rad koji su prethodno ocijenili nastavnici, a koristi se za procjenu kvalitete studentskih ocjena i osiguranje pravedne raspodjele bodova. Studenti nisu unaprijed znali koji je od radova kalibracijski.

U okviru ove laboratorijske vježbe bilo je 10 različitih zadataka, svaki s varijabilnim brojem pitanja u upitnicima. Na temelju tih pitanja studenti su međusobno ocjenjivali radove svojih kolega.

Većina pitanja u upitnicima odnosila se na izgled ER i relacijskog modela te ispravnost definicija entiteta, atributa i veza između entiteta. Manji dio pitanja odnosio se na ispravnost definiranih ograničenja u SQL skripti. Zadnje pitanje u svakom upitniku odnosilo se na sveukupan dojam rada. Sva pitanja vrednovana su ocjenama od 1 do 5 prema sljedećoj skali:

1 – potpuno neispravno

2 – uglavnom neispravno (s većim pogreškama)

3 – napola ispravno

4 – uglavnom ispravno (s manjim pogreškama)

5 – potpuno ispravno

Na raspolaganju za analizu je podatkovni skup koji sadrži ocjene studenata za pet radova koje su trebali vrednovati. *Slika 5.1* prikazuje spomenuti podatkovni skup. Stupac „jobs“ označava jedinstvene identifikatore vrednovanih radova, pri čemu je onaj s negativnim identifikatorom kalibracijski. Stupac „id_student“ označava jedinstveni identifikator studenta koji je ocjenjivao te radove. Stupci od „q1_1“ do „q20_1“ označavaju ocjene koje je student s identifikatorom „id_student“ dao prvom radu iz liste „jobs“. Na isti način su zabilježene i ocjene za ostale radove iz liste „jobs“. Osim prikazanog, dostupni su i podatci o ocjenama nastavnika za taj isti kalibracijski test koji je ocijenio određeni student. Ti podatci će poslužiti za procjenu koliko dobro je određeni student vrednovao radove svojih vršnjaka što će biti opisano u poglavlju **9**.

jobs	id_student	q1_1	q2_1	q3_1	q4_1	q5_1	q6_1	q7_1	q8_1	q9_1	q10_1	q11_1	q12_1	q13_1	q14_1	q15_1	q16_1	q17_1	q18_1	q19_1	q20_1
{322907,-42491,323148,323211,323306}	5754	5	5	5	5	5	5	5	5	5	5	5	2	5	5	5	5	4	5	5	5
{323070,322843,-42466,322814,323209}	5346	5	5	5	5	5	3	3	3	5	5	5	5	3	3	4	5	5			
{322845,-42459,323149,323230,323049}	3288	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
{323217,-42459,323047,323309,323189}	5323	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	3	5	5	5
{-42459,323125,323059,323049,322953}	5335	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	3	1	5	5
{323214,322870,322845,-42459,323217}	5851	5	5	5	5	2	5	5	1	1	5	5	5	5	5	5	5	3	2	2	4
{322990,323049,-42459,323271,323119}	5679	5	5	5	5	5	5	5	5	4	5	4	5	5	5	5	5	4	2	2	4
{323100,323214,-42459,323059,323284}	5706	5	5	5	5	3	5	5	5	2	4	5	5	5	5	5	5	3	4	5	4
{-42459,323256,323280,322895,322984}	5621	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5
{322953,322990,-42459,322870,322845}	5704	5	5	5	5	5	5	5	3	3	5	4	5	5	4	4	4	5	1	1	4
{-42459,322895,329783,323241,323047}	5673	5	5	5	4	5	5	5	5	5	5	5	5	4	4	5	4	4	5	5	4
{323047,322845,322984,323201,-42459}	5463	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
{323249,322895,323143,323256,-42459}	6101	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	3	5	5	5
{-42459,323271,322939,329783,323142}	3344	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
{322956,323100,323189,-42459,322826}	3713	5	4	5	5	5	5	5	4	5	1	3	5	3	5	5	5	5	1	5	4
{323059,-42459,322974,323066,323129}	3493	5	5	5	4	5	5	4	4	4	5	4	5	5	5	5	5	3	5	5	5
{323284,322826,323249,322848,-42459}	5800	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
{323149,322953,323025,322990,-42459}	6131	4	4	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	3	3	5
{322950,323179,-42459,323142,323143}	5367	5	5	5	5	5	5	5	5	5	5	5	5	5	3	5	5	5	5	5	5
{322892,322956,323179,-42459,323241}	3161	5	5	5	5	5	5	5	5	4	5	5	5	5	4	5	5	3	4	4	4
{323119,323309,-42459,322808,322950}	5825	5	5	5	5	5	5	5	5	3	5	5	5	5	5	5	4	3	1	1	4
{323179,323119,323284,323100,-42459}	5813	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5
{323223,323142,-42459,322892,323179}	5398	5	4	5	5	5	5	5	4	2	5	5	5	5	5	5	5	2	4	4	5
{-42459,322875,322848,322953,322870}	3322	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5
{-42459,323241,323284,322950,323271}	5834	5	5	5	5	5	5	5	3	3	5	5	5	5	5	5	5	1	1	1	5
{-42459,323066,323256,323201,329783}	5295	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
{322891,323143,-42459,323007,322895}	5683	5	5	4	5	5	5	5	4	2		3	5	5	5	5	5	2	1	5	4
{-42459,323007,323217,322939,323230}	5859	5	5	5	5	4	5	5	5	4	4	5	5	5	5	5	5	5	1	5	5
{323309,322892,323129,323125,-42459}	5608	5	5	4	4	5	4	4	4	1	4	4	4	4	4	3	3	1	1	1	4
{322808,323249,323280,-42459,322956}	5479	5	5	3	5	5	5	5	3	2	5	5	5	4	3	4	4	3	1	1	4
{323280,-42459,323007,322848,323223}	5864	5	5	5	5	5	5	5	5	2	5	5	5	4	4	5	5	4	1	5	4
{322974,-42459,322875,323223,323309}	5782	5	5	4	5	2	5	5	5	1	5	5	5	5	5	5	5	3	1	5	4
{-42459,322939,323066,322875,323249}	5753	5	5	5	5	5	5	5	4	4	5	5	5	5	5	5	5	5	1	5	4
{-42459,323142,323241,323125,322974}	5374	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
{323049,-42459,323230,322990,322892}	3559	5	5	5	5	3	5	5	5	2	3	3	5	5	5	5	5	2	5	5	3
{322984,-42459,323129,323047,322875}	5743	5	5	3	5	5	5	5	5	5	5	5	5	5	4	5	5	5	1	5	4
{323025,323189,323217,-42459,323280}	5513	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
{323189,-42459,322808,323214,323066}	3246	5	5	2	5	5	5	5	5	1	4	5	5	5	4	5	5	1	1	1	4
{323143,323025,322974,-42459,323149}	5593	4	4	5	4	4	5	5	5	4	5	5	5	5	5	2	2	3	1	1	5
{322848,322984,322891,-42459,323214}	5268	5	5	3	5	5	4	3	3	2	5	5	5	5	5	5	5	3	5	5	5
{323201,323271,322826,322891,-42459}	5520	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	4
{-42459,323129,323223,323100,323059}	5786	5	5	4	5	5	5	5	5	3	5	5	5	5	5	5	5	5	1	5	4

Slika 5.1 - Uzorak podatkovnog skupa s ocjenama vršnjačkog ocjenjivanja

Radove jedne grupe studenata ocijenila su i tri različita nastavnika. Tako su za te studente dostupni podatci o ocjenama od četiri druga studenta i tri nastavnika. Zabilježene su i prosječne ocjene koje su nastavnici ta davali za svaki rad, što se vidi na *Slici 5.2*. Stupac „id_job“ označava jedinstveni identifikator studentskog rada koji se ocjenjuje. Stupci od „q1_t1“ do „q20_t1“ te „avg_t1“ označavaju ocjene koje je prvi nastavnik dao određenom studentskom radu te prosječnu ocjenu tog nastavnika za taj rad. Na isti je način zabilježeno i vrednovanje preostala dva nastavnika.

id_job	q1_t1	q2_t1	q3_t1	q4_t1	q5_t1	q6_t1	q7_t1	q8_t1	q9_t1	q10_t1	q11_t1	q12_t1	q13_t1	q14_t1	q15_t1	q16_t1	q17_t1	q18_t1	q19_t1	q20_t1	avg_t1
322816	5	4	5	5	1	3	5	3	5	5	5	1	1	4	3	5	3	5	2	3	3.65
322827	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322828	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322830	5	2	1	5	3	3	5	1	3	1	5	5	4	2	3	1	1	1	1	2	2.7
322838	5	2	5	5	3	3	5	3	5	5	5	5	1	4	4	5	4	5	4	4	4.1
322864	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	2	5	4.8
322869	5	4	3	5	3	3	4	2	5	3	5	5	3	3	3	5	5	5	2	3	3.8
322874	5	4	5	5	1	4	5	5	5	5	5	1	1	5	4	5	5	5	5	4	4.2
322883	5	5	5	5	3	2	5	1	5	5	5	5	2	3	3	5	3	5	1	4	3.85
322912	5	5	5	5	1	2	5	3	5	5	5	1	1	3	3	5	5	5	1	3	3.65
322929	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322930	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	3	5	4.85
322946	5	5	5	5	5	4	5	5	5	5	5	5	5	4	5	5	4	5	5	5	4.85
322952	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4.95
322962	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	3	5	4.9
322964	4	5	5	5	3	3	5	4	5	5	5	5	4	4	4	5	4	5	1	4	4.25
322965	5	4	5	5	1	3	5	4	5	5	5	1	1	5	4	5	4	5	1	3	3.8
322980	4	5	5	4	1	3	5	4	5	5	5	1	1	4	3	5	4	5	3	3	3.75
322987	5	5	5	5	5	3	5	4	5	5	5	5	4	4	3	5	5	3	3	4	4.4
323010	4	4	5	5	3	3	5	4	5	5	4	4	4	4	3	5	5	4	3	4	4.15
323011	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	3	5	3	5	4.75
323018	5	5	5	5	3	3	5	4	5	5	5	5	4	4	4	5	5	5	3	4	4.45
323046	5	5	5	5	3	3	3	1	5	5	2	5	5	4	3	5	1	1	1	3	3.5
323048	5	5	5	5	3	2	5	4	5	5	5	5	4	4	4	4	5	5	2	4	4.3
323058	4	3	5	4	3	3	4	4	5	5	1	4	4	4	3	5	4	5	2	3	3.75
323060	5	5	5	5	5	3	5	4	5	5	5	5	5	4	5	4	5	5	1	4	4.5
323074	5	4	3	5	5	2	5	4	3	5	5	5	1	4	4	5	4	5	1	3	3.9
323075	5	5	5	5	3	3	5	4	5	5	5	5	3	5	4	4	1	1	1	3	3.85
323076	4	5	5	5	5	5	5	3	5	5	5	5	5	2	5	5	3	5	2	4	4.4
323079	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	1	5	4.7
323083	5	3	3	5	5	4	5	2	5	3	5	5	5	5	4	5	3	5	3	4	4.2
323086	4	5	5	5	5	3	5	4	5	5	5	5	3	3	5	5	5	3	4	4	4.4
323088	5	3	5	5	2	2	4	1	3	3	4	4	3	3	1	1	1	1	1	3	2.75

Slika 5.2 - Uzorak podatkovnog skupa s ocjenama nastavnika

Nadalje, dani su podatci o ukupnom broju bodova koje je svaki student dobio za svoj rad (od maksimalno pet bodova) te o broju bodova koje je dobio za dio zadatka koji se odnosi na vrednovanje radova svojih kolega (također od maksimalno pet bodova). Bodove su studentima dodijelili nastavnici koristeći svoj algoritam.

Ovi podatci omogućit će detaljnu analizu i usporedbu ocjenjivanja između velikih jezičnih modela, studenata i nastavnika.

5.2. Metodologija

Prilikom prikupljanja podataka za analizu korišteni su veliki jezični modeli ChatGPT (korištena verzija ChatGPT 4.0) i Bard/Gemini (korištena verzija Gemini Plus). Svakom modelu je poslan studentski rad u obliku ER modela, relacijskog modela i SQL skripte, kao i tekst zadatka i kriterij ocjenjivanja. Osim ocjena, od modela se tražilo da uz svaki rad napišu kratku povratnu informaciju u kojoj će objasniti svoje ocjenjivanje i ostaviti generalni dojam rada. Primjer jednog zahtjeva poslanog LLM-u prikazan je na *Slika 5.3*, a primjer odgovora ChatGPT-a prikazan je na *Slika 5.*4.

Možeš li mi molim te ocijeniti ovo studentsko rješenje laboratorijske vježbe iz predmeta Baze podataka sa druge godine FER-a, smjer računarstvo za ovaj zadatak:

{tekst zadatka}

Rad vrednuj po ovim kriterijima:

{priloženi kriteriji ocjenjivanja zadatka}

Pogledaj ER model, relacijski model i sql skriptu koju je student predao:

{priloženi ER model, relacijski model i sql skripta}

Slika 5.3 – Primjer zahtjeva poslanog LLM-u

1. Entitet "Izdavač" - 5
2. Entitet "Časopis" - 5
3. Entitet "Perioda" - 5
4. Entitet "Znanstveno područje" - 5
5. Identifikacijski slabi entitet "Izdanje" - 5
6. Identifikacijski slabi entitet "Članak" - 5
7. Entitet "Autor" - 5
8. Refleksivna veza "Referencira" - 4
9. Veza između "Izdavač" i "Časopis" – 5
10. Veza između "Perioda" i "Časopis" – 5
11. Veza između "Znanstveno područje" i "Časopis" – 5
12. Veza između "Izdanje" i "Časopis" – 5
13. Veza između "Izdanje" i "Članak" – 5
14. Veza između "Autor" i "Članak" – 5
15. Relacijske sheme – 5
16. Tipovi i duljine atributa – 5
17. Ograničenje chkGodinaOsnutka – 5
18. Ograničenje chkRbrClanak – 5
19. Ograničenje chkRef – 5 (no details given)
20. Cjelokupni dojam – 5

Komentar: Rješenje izgleda korektno i u skladu s postavljenim zahtjevima zadataka. ER i relacijski modeli su dobro dizajnirani, a SQL skripta sadrži sve potrebne entitete, atribute i integritetska ograničenja.

Slika 5.4 – Primjer odgovora ChatGPT-a

Nakon što su modeli ocijenili radove, njihove su ocjene zabilježene s ciljem usporedbe s ocjenama koje su za iste radove dali i studenti i nastavnici, kao i prosječne ocjene koje je određeni LLM dao za određeni rad, što se vidi na *Slika 5.5*. Stupac „id_job“ predstavlja jedinstveni identifikator studentskog rada koji se ocjenjuje. Stupci od „q1_cgpt“ do „q20_cgpt“ označavaju ocjene koje je ChatGPT dodijelio tom studentskom radu, dok stupac „avg_cgpt“ označava njegovu prosječnu ocjenu za taj rad. Na isti način su zabilježene i ocjene Bard-a.

id_job	q1_cgpt	q2_cgpt	q3_cgpt	q4_cgpt	q5_cgpt	q6_cgpt	q7_cgpt	q8_cgpt	q9_cgpt	q10_cgpt	q11_cgpt	q12_cgpt	q13_cgpt	q14_cgpt	q15_cgpt	q16_cgpt	q17_cgpt	q18_cgpt	q19_cgpt	q20_cgpt	avg_cgpt
322827	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322828	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322838	5	5	5	5	3	3	5	4	5	5	5	3	3	5	4	5	5	5	3	4	4.35
322864	5	5	5	5	4	4	5	4	5	5	5	5	5	5	5	5	5	5	4	4	4.75
322869	5	5	5	5	4	4	5	4	5	4	5	4	4	5	3	5	5	5	3	3	4.4
322874	5	4	5	5	3	3	5	4	4	4	4	3	3	5	4	5	4	5	4	4	4.15
322883	5	4	5	5	4	3	5	5	5	5	5	4	4	5	4	5	4	5	5	4	4.55
322912	4	4	5	5	3	2	5	4	5	5	4	3	2	4	3	5	3	5	5	3	3.95
322929	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	2	5	4.8
322930	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322946	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322952	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
322962	5	5	5	5	5	5	5	5	5	5	5	5	4	5	4	5	5	5	2	5	4.75
322964	4	3	5	5	2	2	5	1	4	5	3	3	5	5	5	5	1	5	2	3	3.65
322987	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	4	4.55
323010	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	1	1	3	4.45
323018	5	4	5	5	4	4	5	3	5	5	5	4	4	5	5	5	5	5	2	4	4.45
323048	5	4	5	5	3	3	5	3	5	5	5	4	4	5	5	5	5	5	2	4	4.35
323058	5	4	5	5	4	3	5	5	5	5	5	5	5	5	5	5	5	5	4	4	4.7
323060	5	5	5	5	5	2	5	2	5	5	5	5	5	5	5	5	5	5	2	4	4.5
323074	5	5	5	5	4	2	5	3	5	5	5	5	2	5	4	5	4	5	2	4	4.25
323076	5	5	5	4	4	4	5	4	5	5	5	5	5	5	4	5	5	5	4	3	4.6
323079	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	4	4.75
323083	5	5	5	5	5	3	5	4	5	5	5	5	4	5	4	5	3	5	3	4	4.5
323086	5	5	5	5	4	3	5	5	5	5	5	4	3	5	5	5	4	5	3	5	4.55
323094	5	5	5	5	5	2	5	5	5	5	5	5	2	5	5	5	5	5	5	5	4.7
323112	5	5	5	5	5	4	5	4	5	5	5	5	5	5	4	5	5	5	4	4	4.75
323123	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	2	5	4.85
323134	5	5	5	5	2	2	5	5	5	5	5	2	2	5	4	5	5	5	2	4	4.15
323145	5	5	5	5	5	3	5	5	5	5	5	5	4	5	5	5	5	5	5	5	4.85
323154	5	4	5	5	2	2	5	1	5	5	4	2	2	4	3	5	5	5	1	3	3.65
323159	5	5	5	5	5	3	5	3	5	5	5	5	5	5	5	5	5	5	5	4	4.75
323162	5	5	5	5	4	4	5	4	5	5	4	5	5	5	5	4	5	5	1	4	4.5
323170	4	4	3	5	3	3	5	5	5	3	5	4	4	5	4	5	5	3	5	4	4.2

Slika 5.5 – Uzorak podatkovnog skupa s ocjenama LLM-ova

Ono što je zanimljivo je da Bard nije mogao ocijeniti svih 20 pitanja odjednom; vraćao je povratne informacije samo za prvih nekoliko pitanja i potom bi stao. Zbog toga su Bardu slana pitanja u skupinama od po pet, pri čemu su mu svaki put ponovo dostavljeni svi potrebni materijali za ocjenjivanje (studentsko rješenje, tekst zadatka i kriterije ocjenjivanja). S druge strane, ChatGPT nije imao takvih problema. Svaki put je vratio povratnu informaciju za svih 20 pitanja odjednom.

6. Interpretacija rezultata prikupljenih od LLM-ova

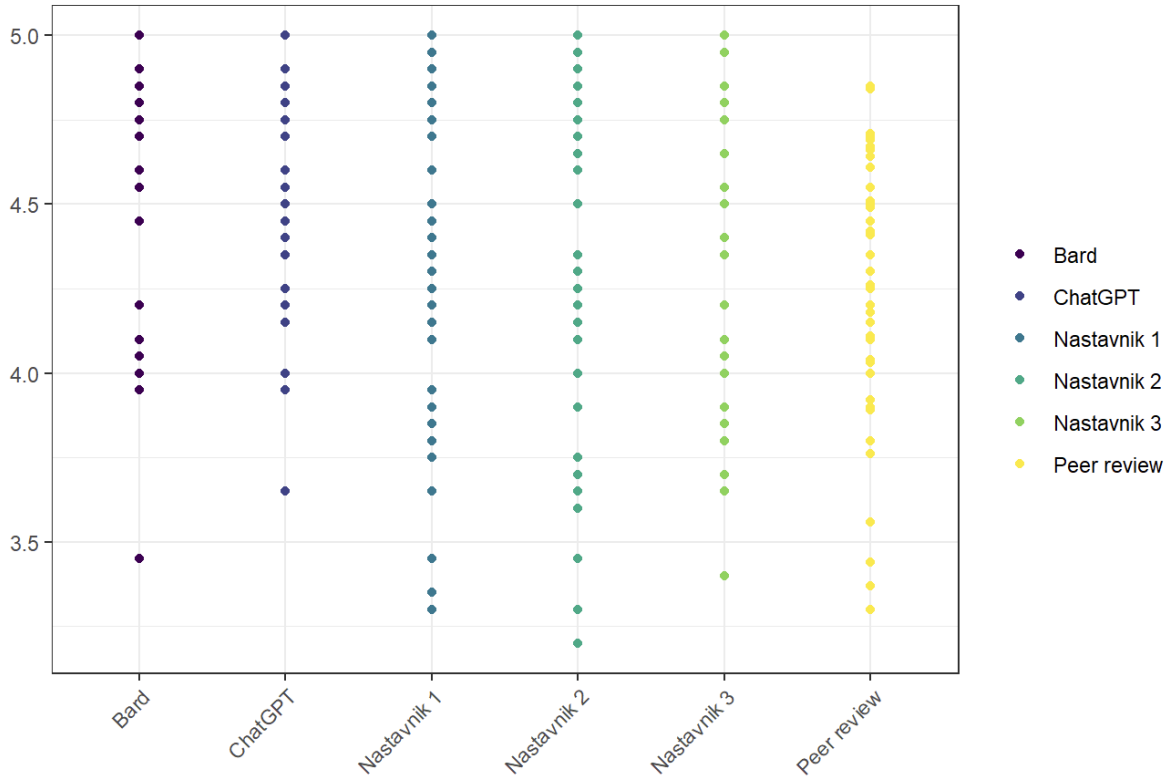
U nastavku je opisana interpretacija rezultata provedene analize podataka. Promatrajući različite numeričke vrijednosti i frekvencije dodjeljivanja pojedinih ocjena, napravljena je usporedba kako su različiti ocjenjivači vrednovali iste radove te usporedba ocjenjivanja LLM-ova i nastavnika po pitanjima na primjeru jednog nasumično izdvojenog studentskog rada.

6.1. Usporedba prosječnih ocjena

Graf prikazan na *Slici 6.1* vizualizira prosječne ocjene koje su dodijeljene studentskim radovima od strane pet različitih ocjenjivača: Bard-a, ChatGPT-a, drugih studenata kroz peer review način ocjenjivanja te tri različita nastavnika označena kao „Nastavnik 1“, „Nastavnik 2“ i „Nastavnik 3“. Svaka točka na grafu predstavlja prosječnu ocjenu dodijeljenu jednom studentskom radu.

Najširu raspodjelu ocjena pokazuju nastavnici. Peer review ocjenjivanje, koje provode sami studenti, teži prema nižim ocjenama. Vrlo malo radova je ocijenjeno prosječnom ocjenom višom od 4.5. Ovo može biti rezultat nekoliko faktora, uključujući nedostatak iskustva u ocjenjivanju te strah od gubitka bodova u drugom dijelu laboratorijske vježbe u slučaju da je njihovo ocjenjivanje bilo preblago.

Bard i ChatGPT pokazuju najmanju varijabilnost u ocjenjivanju, s ocjenama koje su većinom grupirane oko viših prosječnih vrijednosti.



Slika 6.1 - Usporedba prosječnih ocjena LLM-ova, studenata i nastavnika

6.2. Box-plot vizualizacija

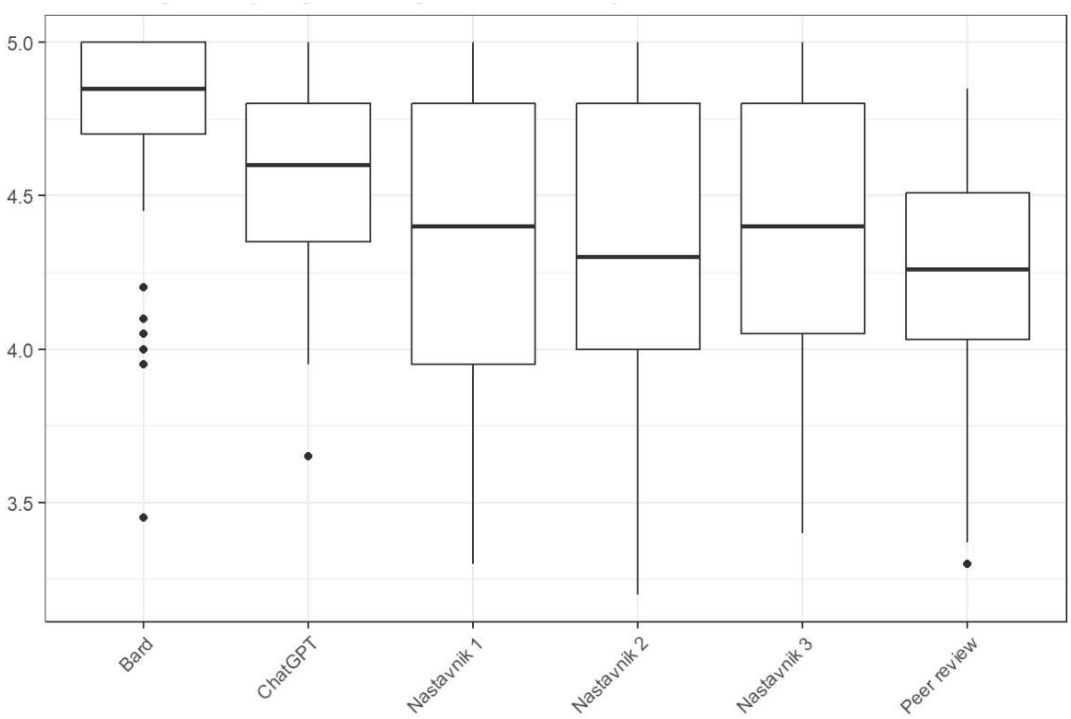
Box-plot grafovi su korisni za vizualizaciju raspodjele ocjena jer prikazuju medijane, interkvartilne raspone i stršeće vrijednosti (*engl. outliers*). Na grafu prikazanom na *Slika 6.2* nalazi se pet box-plotova koji predstavljaju ocjene dodijeljene pojedinim studentskim radovima od strane Bard-a, ChatGPT-a, peer review ocjene te ocjene tri različita nastavnika.

Medijani ocjena koje su dodijelili ChatGPT i Bard nalaze se između 4.5 i 5.0, što ukazuje na tendenciju ovih ocjenjivača da dodjeljuju više ocjene. S druge strane, medijan peer review ocjena je najniži, iz čega se može zaključiti da su studenti bili najstroži u svojoj evaluaciji.

Box-plot Bard-ovih ocjena ima relativno uži interkvartilni raspon u odnosu na ostale ocjenjivače, što ukazuje na manju varijabilnost u ocjenjivanju.

Bard, ChatGPT i peer review ocjene pokazuju prisutnost stršećih vrijednosti, što je svojevrsni pokazatelj nekonzistentnosti u vrednovanju. Ova nekonzistentnost može proizlaziti iz subjektivnih pristupa tih ocjenjivača, što nije slučaj kod vrednovanja

nastavnika. Ocjene nastavnika pokazuju najmanje varijacije, odnosno nastavnici među sobom imaju puno ujednačeniji pristup ocjenjivanju.



Slika 6.2 - Box-Plot grafovi prosječnih ocjena LLM-ova, studenata i nastavnika

6.3. Numeričke vrijednosti

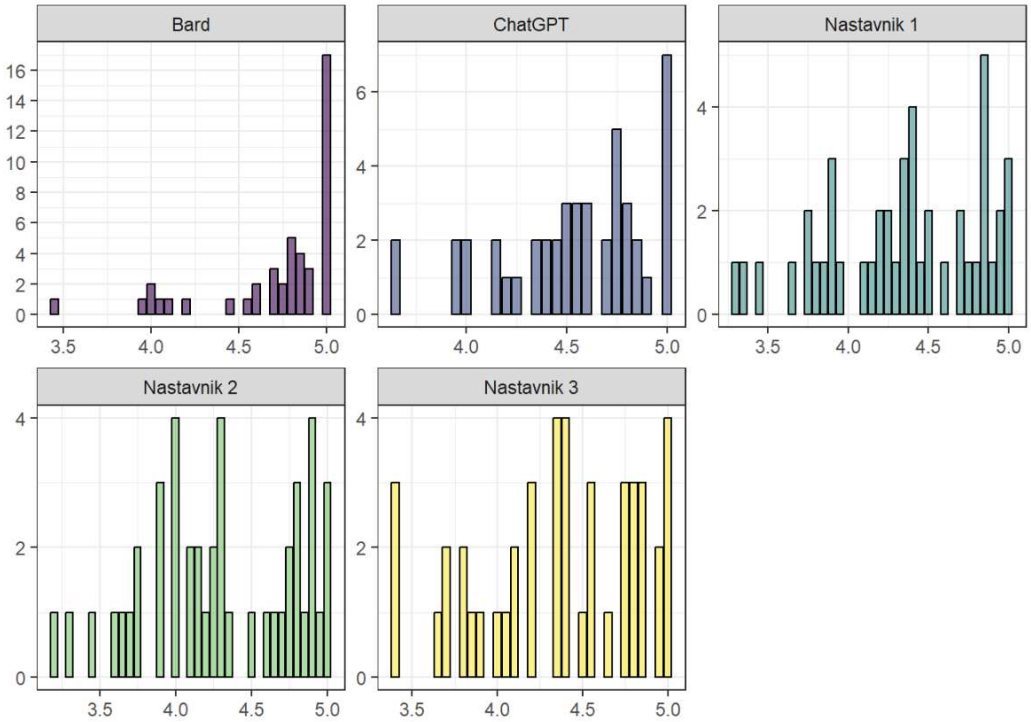
Numerički podaci iz *Tablica 6.1* prikazuju prosječnu ocjenu koju je pojedini ocjenjivač dodijelio studentskim radovima te medijan ocjena dodijeljenih od strane svakog ocjenjivača. Ovi podatci potvrđuju ono što smo vizualno primijetili na grafovima, Bard i ChatGPT su davali generalno više ocjene, no ocjene ChatGPT-a manje odstupaju od peer review i nastavničkih ocjena. Zanimljivo je da i ova statistika pokazuje kako su peer review ocjene niže i od nastavničkih ocjena, što podupire prvotnu hipotezu da studenti u prosijeku radove svojih kolega ocjenjuju strože no što to rade nastavnici.

OCJENJIVAČ	PROSJEČNA OCJENA	MEDIJAN OCJENA
Bard	4.73	4.85
ChatGPT	4.54	4.60
Peer review	4.25	4.26
Nastavnik 1	4.35	4.40
Nastavnik 2	4.30	4.30
Nastavnik 3	4.36	4.40

Tablica 6.1 - Ukupan prosjek i medijan ocjena različitih ocjenjivača

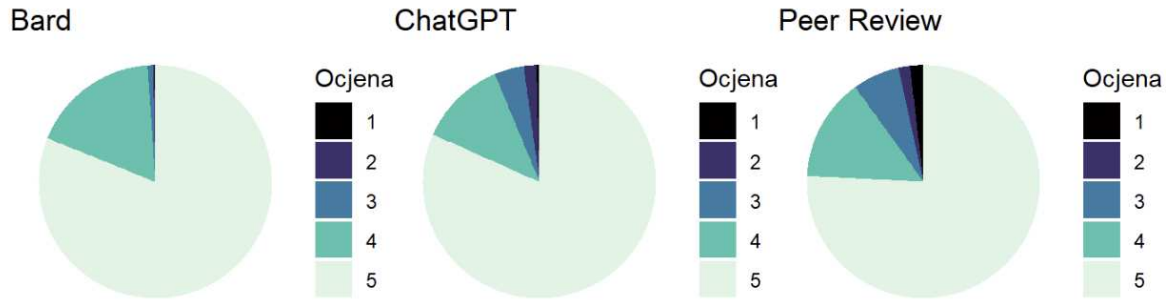
6.4. Frekvencije ocjena

Na *Slika 6.3* može se vidjeti vizualna interpretacija frekvencija prosječnih ocjena dodijeljenih za studentske radove od strane dva LLM-a te tri nastavnika. Na x-osi nalaze se vrijednosti prosječnih ocjena, dok se na y-osi nalaze frekvencije svake od ocjena za pojedine ispravljače. Vidi se da Bard najčešće radovima dodjeljuje maksimalan broj bodova (sedamnaest puta zabilježena prosječna ocjena 5.0), dok je kod nastavnika i ChatGPT-a prisutan puno širi spektar ocjena.



Slika 6.3 - Prikaz frekvencija prosječnih ocjena LLM-ova i nastavnika

Na grafovima prikazanim na *Slika 6.4* mogu se vidjeti frekvencije pojedinih ocjena kojima su LLM-ovi i peer review ocjenjivači vrednovali zadatke. Vidi se kako su sva tri ocjenjivača najčešće vrednovali zadatke s ocjenom 5, no raspodjela ostalih ocjena studenata i ChatGPT-a ukazuju na njihovu veću objektivnost u ocjenjivanju u odnosu na Bard, koji je u preko 90% slučajeva dodijelio ocjene 4 ili 5.



Slika 6.4 - Prikaz frekvencija ocjena LLM-ova i studenata

6.5. Usporedba po pitanjima

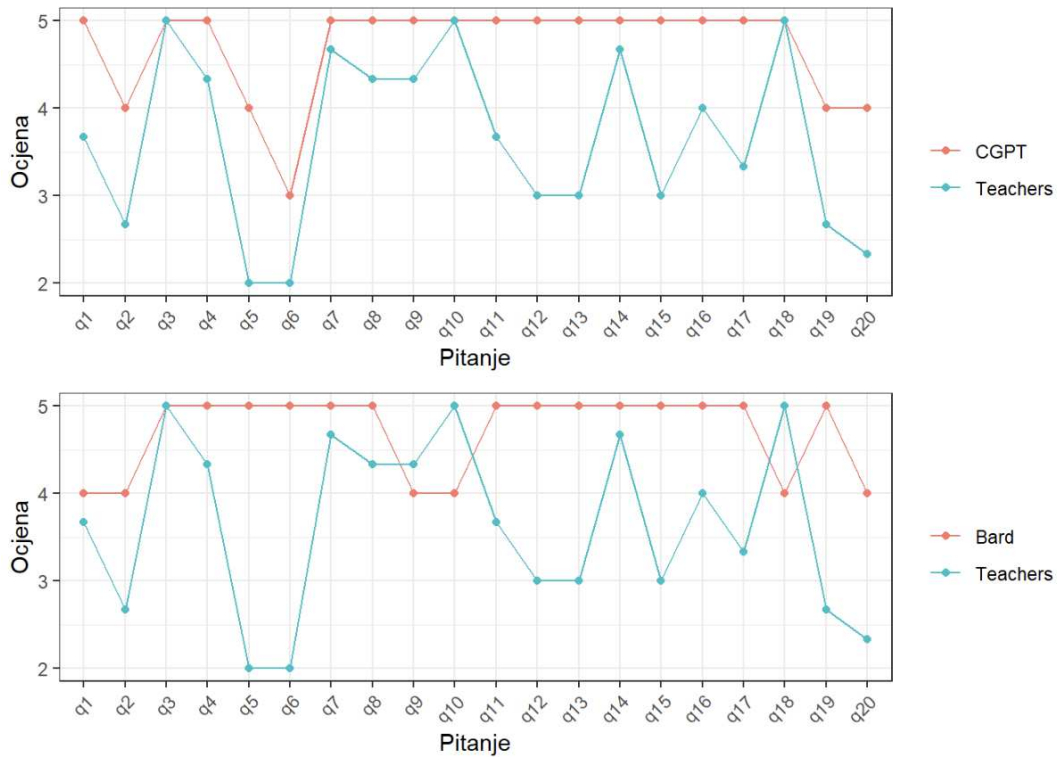
Na grafu prikazanom na *Slika 6.5* vidljiva je usporedba ocjena nastavnika i LLM-ova po pitanjima za nasumično odabranog studenta. Kao ocjena nastavnika uzeta je aritmetička sredina ocjena sva tri nastavnika za svako pitanje rada tog studenta.

Nakon analize ovakvih grafova za ostale studente, može se zaključiti da ChatGPT i Bard za pojedina pitanja imaju velika odstupanja u odnosu na nastavnike. Npr. u slučajevima gdje nastavnici vrednuju neko pitanje ocjenom 2 ili 3, LLM-ovi ga vrednuju ocjenama 4 ili 5 što bi značilo da je po njima taj dio zadatka odrađen besprijekorno, a to nije slučaj.

Najveća odstupanja su se pokazala na 19.pitanju, koje se odnosi na provjeru ispravnosti implementiranog ograničenja u SQL skripti. Kod nekih studenata implementacija danog ograničenja nije bila niti prisutna, no LLM-ovi su zadatak nagradili maksimalnim brojem bodova.

Osim toga, veća odstupanja se vide i kod 5. i 6. pitanja koja se odnose na provjeru je li određeni entitet u ER dijagramu označen kao slabi entitet.

12. i 13. pitanje su također dosta loše ocijenjeni od strane LLM-ova, a ta pitanja se odnose na provjeru je li određena veza u ER modelu označena kao identificirajuća.



Slika 6.5 - Prikaz ocjena za pojedinačnog studenta po pitanjima

7. Analiza komentara

U ovom poglavlju je predstavljena usporedba komentara koje su za određene radove ostavili LLM-ovi s komentarima koje su za iste te radove ostavili studenti. Također, prikazano je kako se komentari i ocjenjivanje LLM-ova slažu s ocjenama koje su davali nastavnici, što će dati uvid u to koliko su LLM-ovi bili sposobni uočavati određene greške u radu i, ako su ih uočili, jesu li ih pravedno vrednovali.

7.1. Analiza komentara za odlično riješene radove

U prvom dijelu je napravljena analiza komentara koje su LLM-ovi davali za radove koji su, prema ocjenama nastavnika, dobro riješeni. Analiza se temelji na dva studentska rada koja su sva tri nastavnika vrednovala s prosječnom ocjenom 5.0.

Tablica 7.1 prikazuje komentare i ocjenu ChatGPT-a i Bard-a za prvi takav rad. Vidi se da su primijetili da je predani rad u potpunosti ispravan te su adekvatno dodijelili bodove i komentare.

	ChatGPT	**Bard**
Komentar	„Tablice su definirane s pravilnim tipovima podataka i duljinama, a ograničenja su ispravno postavljena. Implementacija veza između entiteta sugerira da je dobro znanje o strukturi i odnosima unutar baze.“	„Rješenje je u potpunosti ispravno i nema grešaka. ER model je dobro nacrtan. Relacijski model je ispravno implementiran. Ograničenja su ispravno definirana i tipovi podataka su ispravno postavljeni.“
Prosječna ocjena	5.0	5.0

Tablica 7.1 - Komentari i prosječne ocjene LLM-ova za prvi rad

U *Tablica 7.2* vide se komentari koje su studenti ostavili taj rad. Studenti su također ovom radu dodijelili maksimalan broj bodova za sva pitanja te ostavili komentare koji podržavaju njihove ocjene.

	Student 1	Student 2
Komentar	„Nemam negativnih komentara, Ovo je savršeno napravljen model u apsolutno svim područjima“	„Nisam našao niti jednu grešku, perfect.”

Tablica 7.2 - Komentari studenata za prvi rad

S druge strane, *Tablica 7.3* prikazuje komentare i ocjene LLM-ova za drugi takav rad iz čega se može primijetiti kako su odlučili za neka pitanja dati broj bodova manji od maksimalnog te je zato i prosječna ocjena manja od 5.0. Pri tome, nisu u komentaru obrazložili zašto su odlučili uzeti bodove za neka pitanja.

	ChatGPT	Bard
Komentar	„Oba dijagrama i SQL kôd kvalitetno izrađeni“	„Dijagram je dobro nacrtan i pravilno prikazuje entitete, atribute i relacije u bazi podataka.“
Prosječna ocjena	4.85	4.8

Tablica 7.3 - Komentari i prosječne ocjene LLM-ova za drugi rad

Kao ni LLM-ovi, ni studenti nisu ocijenili ovaj rad s maksimalnim brojem bodova, odnosno prosječnom ocjenom 5.0. No, za razliku od LLM-ova, komentari studenata *(Tablica 7.4)* su puno konstruktivniji i objašnjavaju konkretnu grešku koju su našli u radu. Zanimljivo je kako su nastavnici, bez obzira na malu grešku u radu, i dalje dodijelili maksimalan broj bodova, dok su studenti za to oduzimali bodove. To objašnjava rezultate iz poglavlja **6.** koji pokazuju kako su studenti generalno davali niže ocjene jedni drugima u odnosu na nastavnike.

	Student 1	Student 2
Komentar	„Sve je više manje jednako službenim rješenjima. Jedino u integritetskom ograničenju chkRef se za negaciju koristi "!" umjesto NOT”	„Pogreška u chkRef -> u SQL se ne koristi operator '!' Iako je sintaksa netočna, ideja rješenja je dobra pa ocjenjujem s uglavnom ispravno”

Tablica 7.4 - Komentari studenata za drugi rad

Analizirajući komentare LLM-ova za ostale izvrsno riješene studentske radove poput ova dva, primijećeno je kako je u puno više slučajeva situacija slična kao ona u *Tablica 7.3* kada LLM-ovi studentima ne dodijele zasluženi maksimalni broj bodova bez da opravdaju ili obrazlože svoju odluku.

7.2. Analiza komentara za radove s dosta grešaka

U drugom dijelu su analizirani radovi koji su ispodprosječno riješeni, odnosno imaju dosta grešaka u implementaciji rješenja. U obzir su uzeta dva rada za koje su nastavnici dodijelili prosječnu ocjenu 3.4.

U *Tablica 7.5* vidi se prikaz komentara za jedan ovakav rad. ChatGPT je dao prilično visoku prosječnu ocjenu u odnosu na nastavnike. Također, nije primijetio greške u implementaciji ograničenja u SQL dijelu rješenja te je sva pitanja vezana uz to ocijenio s ocjenom 5, dok su nastavnici za ta pitanja dodijelili minimalan broj bodova. Osim toga, neka pitanja za koja su u potpunosti točna je ocijenio ocjenama 3 ili 4.

Iz Bard-ove prosječne ocjene koja je bliža nastavničkoj, a i popratnog komentara, vidi se da je on primijetio kako postoje nedostatci u implementaciji ograničenja, no on je za ta pitanja dodijelio ocjenu 4, dok su pitanja trebala biti ocijenjena s ocjenom 1 uzevši u obzir da potrebna ograničenja nisu uopće bila implementirana. Kao i ChatGPT, Bard je također neka u potpunosti točna pitanja vrednovao ocjenom 4.

	ChatGPT	**Bard**
Komentar	„Potrebno je detaljnije obratiti pažnju na definiciju i implementaciju integritetskih ograničenja te poboljšati preciznost u sintaksi SQL naredbi.“	„Postoje greške u pojedinim atributima i optimizaciji ograničenja.“
Prosječna ocjena	4.2	3.95

Tablica 7.5 - Komentari i prosječne ocjene LLM-ova za četvrti rad

Iz *Tablica 7.6* jasno je vidljivo da su studenti, suprotno LLM-ovima, vrlo precizni prilikom ostavljanja komentara, točno navodeći greške zbog kojih su oduzeli bodove. Zanimljivo je da ne primjećuju i ne percipiraju svi studenti iste stvari kao greške.

	Student 1	Student 2
Komentar	„Jedini značajniji problem rješenja je nepostojanje entiteta izdanje“	„Nema dijela gdje se mogu dodatno penalizirati neka integritetska ograničenja - npr. nazCasopis je UNIQUE, no nije postavljen kao NOT NULL (nije pritom primary key i time se ne ograničava od potencijalne NULL vrijednosti), tako da može poprimiti NULL vrijednost. Tako da sam to penalizirala na 16. pitanju. Kod chkGodinaOsnutka skinuto 2 boda jer nedostaje CHECK te jedna zagrada što predstavlja syntax error pri izvođenju SQL naredbe. Kod chkRef ne odgovara upit traženom - nedostaje dodatan atribut prema kojem se povjerava ograničenje te ne ide AND već OR.„

Tablica 7.6 - Komentari studenata za drugi rad

U drugom ovakvom radu, vidi se da je prosječna ocjena ChatGPT-a vrlo bliska nastavničkoj *(Tablica 7.7)*. Za dva pitanja koja su u potpunosti krivo implementirana u ovom radu je primijetio da su krivo implementirani te im dao ocjenu u skladu s tim, a svoju odluku je potkrijepio u komentaru.

Bard je, s druge strane, u potpunosti krivo ocijenio ovaj rad. Prosječna ocjena jako odstupa od prosječne ocjene nastavnika. Iz komentara se može vidjeti kako Bard uočava da SQL kod nije točan u potpunosti no sva pitanja vezana uz to je ocijenio s maksimalnim brojem bodova.

Zanimljivo je kako i jedan i drugi LLM koriste izraze poput „potrebna je dodatna provjera“ za određeni dio rada. Ovakav pristup sugerira da se izbjegava direktno navođenje grešaka ili nedostataka, već se koriste općeniti izrazi.

	ChatGPT	Bard
Komentar	„Rješenje prikazuje solidno razumijevanje ER modeliranja i relacijskih baza podataka, uz nekoliko manjih nedostataka koji ne umanjuju funkcionalnost baze. Primarni i jedinstveni ključevi su većinom ispravno definirani, no ključna veza za „Članak" bi mogla biti bolje integrirana s entitetom "Izdanje". Također, detalji implementacije ograničenja zahtijevaju dodatnu provjeru.“	„Rad pokazuje solidno razumijevanje koncepata modeliranja baza podataka. S provjerom SQL koda i malo više detalja u dokumentaciji, model bi bio izvrstan.
Prosječna ocjena	3.65	4.7

Tablica 7.7 - Komentari i prosječne ocjene LLM-ova za treći rad

Sve u svemu, može se primijetiti da, bez obzira na to je li rad ispravno riješen ili ima puno grešaka, ocjenjivanje i komentari od strane LLM-ova su nepredvidljivi. U jednom slučaju obave posao poprilično dobro i potkrijepe svoje vrednovanje relativno dobrim komentarom, dok u drugom slučaju njihov komentar uopće ne odražava ocjene koje su dodijelili.

8. Evaluacija ocjenjivanja zadataka otvorenog tipa

Nakon detaljne analize ocjena i komentara koje su pružili LLM-ovi, u ovom poglavlju je ocijenjena njihova efikasnost.

U daljnjoj analizi, umjesto razmatranja ocjena koje je student dobio za pojedino pitanje od tri različita nastavnika zasebno, radi jednostavnosti uzet će se prosjek njihovih ocjena kao jedinstvena ocjena. Na primjer, za prvo pitanje pojedinog studentskog rada uzet će se prosjek tri ocjene koje su mu dodijelila tri nastavnika za to pitanje.

Kako bi se izračunala mjera odstupanja, odnosno pogreške vrednovanja velikih jezičnih modela u odnosu na vrednovanje nastavnika, definirana je funkcija gubitka kao omjer kvadratnih odstupanja pogreške (SSE) i ukupnog kvadratnog odstupanja (SST). Ta mjera pokazuje koliki udio ukupnog kvadratnog odstupanja otpada na odstupanje pogreške modela. Osim toga, izračunat je koeficijent determinacije za velike jezične modele, koji je pokazatelj kolika je proporcija varijacije u zavisnoj varijabli prihvatljiva, odnosno predvidljiva nezavisnom varijablom.

Za početak je potrebno izračunati vrijednosti SSE za ChatGPT i za Bard te vrijednost SST točnih vrijednosti zavisne varijable cijelog podatkovnog skupa. Rezultati tog izračuna prikazani su u *Tablica 8.1*. Može se primijetiti da je SSE ChatGPT-a manji od SSE Bard-a, što znači da su ocjene ChatGPT-a manje odstupale od ocjena nastavnika, odnosno, bio je puno točniji u vrednovanju zadataka nego što je to bio Bard.

Suma kvadratnih odstupanja ocjena ChatGPT-a od ocjena nastavnika (SSE)	630.62
Suma kvadratnih odstupanja ocjena Bard-a od ocjena nastavnika (SSE)	1121.416
Suma kvadratnih odstupanja ocjena nastavnika i njihovog prosjeka	973.582

Tablica 8.1 - Rezultati kvadratnih odstupanja

U *Tablica 8.2* prikazani su iznosi funkcija gubitaka, dok su u *Tablica 8.3* prikazani izračuni koeficijenta determinacije za ChatGPT i Bard. Funkcija gubitka za ChatGPT pokazuje da greške koje čini ChatGPT obuhvaćaju oko 64.77% ukupne varijabilnosti ocjena. To implicira da ChatGPT objašnjava otprilike 35.23% varijabilnosti ocjena. Iako nije besprijekoran, ovaj model ipak pokazuje određenu razinu točnosti.

Za Bard se ne može donijeti isti zaključak. Budući da je funkcija gubitka kod Bard-a veća od 1, to znači da su Bard-ove pogreške veće od ukupne varijabilnosti ocjena nastavnika. S druge strane, to se očituje negativnim koeficijentom determinacije, što znači da model ne samo da ne objašnjava nikakvu varijaciju, već i daje lošije predikcije ocjena nego što bi ih dali nasumično. Drugim riječima, Bard-ova predviđanja su toliko loša da unose novu varijabilnost umjesto da ju objasne.

Funkcija gubitka za ChatGPT	0.6477
Funkcija gubitka za Bard	1.1518

Tablica 8.2 - Rezultati funkcije gubitka

Koeficijent determinacije za ChatGPT	0.3523
Koeficijent determinacije za Bard	-0.1518

Tablica 8.3 – Rezultati koeficijenta determinacije

9. Algoritam za dodjeljivanje bodova na temelju kalibracijskog testa

U svrhu procjene sposobnosti vrednovanja zadataka ovog tipa, predložena je alternativna implementacija algoritma za dodjeljivanje bodova u odnosu na algoritam koji su koristili nastavnici. Algoritam se oslanja na točnost ocjenjivanja kalibracijskog testa. Testiran je na skupu podataka koji sadrži rezultate prethodno spomenute laboratorijske vježbe iz predmeta „Baze podataka“. Osim toga, ovaj algoritam je uspoređen s algoritmom kojeg su nastavnici koristili pri dodjeli stvarnih bodova u okviru laboratorijske vježbe.

Algoritam se sastoji od dva dijela. Prvi dio odnosi se na računanje bodova koje bi pojedini student trebao dobiti za svoje ocjenjivanje, odnosno „peer review“ dio laboratorijske vježbe. Drugi dio algoritma fokusira se na računanje bodova koje bi student trebao dobiti za svoj vlastiti rad.

9.1. Dodjeljivanje bodova za ocjenjivanje

U ovom segmentu se za analizu koristi podatkovni skup koji obuhvaća ocjene koje su studenti dodijelili kalibracijskom testu, uspoređujući ih s „točnim“ ocjenama dodijeljenim istom testu od strane nastavnika.

Tijekom analize, primijećeni su zapisi u podatkovnom skupu s nedostajućim podacima, što ukazuje da su neki studenti propustili ocijeniti radove svojih kolega u peer review dijelu laboratorijske vježbe. Takvi slučajevi su izuzeti iz daljnje analize jer ti uzorci nisu dovoljno reprezentativni i mogli bi narušiti valjanost rezultata i zaključaka. Studentima koji nisu sudjelovali u ocjenjivanju kalibracijskog testa je automatski dodijeljeno nula bodova za ovaj dio vježbe.

Prvo su uspoređene studentove ocjene s točnim ocjenama istih pitanja te su na temelju toga izračunate mjere SSE, SSR i SST. Nakon toga, izračunate su "težine", definirane kao omjer SSR i SST, odnosno koeficijent determinacije. Ovo se pokazalo kao koristan pristup jer koeficijent determinacije pruža uvid u postotak varijacije koji je objašnjen, što nam omogućava procjenu točnosti ocjenjivanja. Zatim se na težine primjenjuje linearna

transformacija na željeni interval. Nakon toga se dobiveni koeficijent množi s maksimalnim brojem bodova (u ovom slučaju pet), čime određujemo broj bodova koje će student dobiti za svoje ocjenjivanje.

Kôd 9.1 prikazuje ključne korake ovog algoritma. Za svakog studenta izračunavaju se mjere SST [1], SSE [2] i SSR [3]. Zatim se izračunavaju koeficijenti determinacije, koji se označavaju listom "weights" [4]. Ti koeficijenti se potom transformiraju na željeni interval koristeći funkciju „linear_transform“ [5]. Nakon toga se iterira kroz podatkovni skup koji sadrži sve studente [6], pronalazi se izračunata težina za svakog studenta [7], te se u liniji [8] izračunava broj bodova koje bi student trebao dobiti za svoje ocjenjivanje kao umnožak transformiranog koeficijenta i maksimalnog mogućeg broja bodova.

```
1  SST = [sum((row - row.mean()) ** 2) for _, row in student_grades.iterrows()]
2  SSE = [sum((row - correct) ** 2) for (_, row), (_, correct) in
           zip(student_grades.iterrows(), correct_grades.iterrows())]
3  SSR = [sst - sse if sst != 0 else None for sst, sse in zip(SST, SSE)]
4  weights = [1 - (sse / sst) for sse, sst in zip(SSE, SST)] # rsquared
5  transformed_weights = linear_transform(weights, target_min, target_max)
6  for idx, row in calibration_data.iterrows():
7          coefficient = abs(transformed_weights[idx])
8          prediction = coefficient * 5
```

Kôd 9.1 - Algoritam za dodjeljivanje bodova za ocjenjivanje

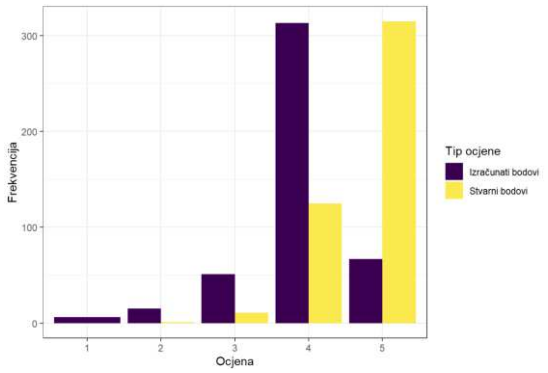
Na koeficijente se primjenjuje linearna transformacija kako bi se ublažili. Originalni koeficijenti su izračunati kao koeficijent determinacije koji će uglavnom poprimiti vrijednosti između 0 i 1 kao što je objašnjeno u poglavlju **2.2**. To znači da će student koji je jako dobro ocijenio rad imati koeficijent [0.9-1], dok će student koji je napravio manji broj pogrešaka imati koeficijent [0.5-0.6]. Kada bi koeficijenti ostali takvi, algoritam bi previše kažnjavao male pogreške i dodjeljivao studentima manje bodova nego što zaslužuju. Zato se na koeficijente primjenjuje linearna transformacija koja će ih preslikati u određeni interval kako bi se pogreške u ocjenjivanju kažnjavale u manjoj mjeri. Također, neki studenti su toliko loše ocijenili kalibracijski test da imaju negativni koeficijent determinacije koji će se onda primjenom linearne transformacije preslikati u neki jako mali pozitivni koeficijent kako bi se studentu adekvatno mogli dodijeliti bodovi.

Za početak se na koeficijente probala primijeniti linearna transformacija na interval [0.6, 1]. Kada se usporede bodovi dobiveni algoritmom sa stvarnim bodovima koje je svaki

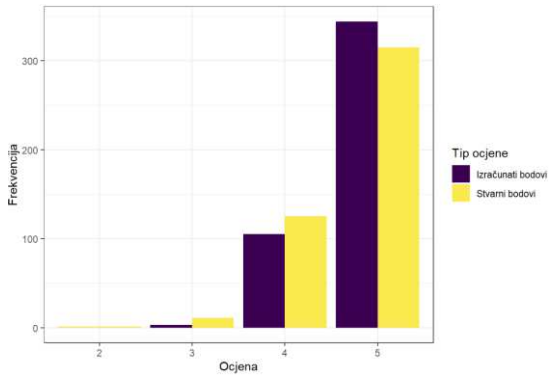
student dobio, dobije se samo 60% podudaranja, što je dosta malo. Analizom se primjećuje kako su koeficijenti i dalje dosta strogi što rezultira i strožim dodjeljivanjem ocjena. To se može vidjeti i na *Slici 9.1* gdje je prikazano kako algoritam u ovom slučaju jako malom broju studenata dodijeli maksimalan broj bodova.

Međutim, kada se smanji raspon intervala unutar kojeg se vrši linearna transformacija koeficijenata na [0.75, 1], postotak podudaranja ocjena se povećava. Daljnjim smanjenjem intervala, na primjer na [0.85, 1], postotak podudaranja raste na 75%. Ipak, ako se još nastavi smanjivati interval, postotak podudaranja će se opet smanjivati, što znači da bi u tom slučaju koeficijenti postali preblagi, što se i vidi na *Slici 9.1* gdje su studenti prilikom korištenja intervala [0.9-1] ovim algoritmom dobili puno više puta maksimalan broj bodova u usporedbi sa stvarnim bodovima.

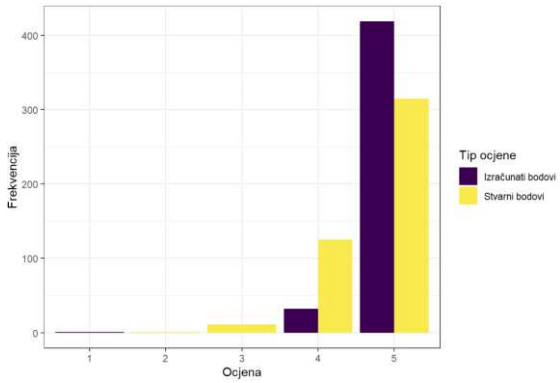
Primjena linearne transformacije na interval [0.6-1]



Primjena linearne transformacije na interval [0.85-1]



Primjena linearne transformacije na interval [0.9-1]



Slika 9.1 - Prikaz frekvencije bodova dodijeljenih predloženim algoritmom primjenom linearnih transformacija različitih intervala na koeficijente i stvarnih bodova

9.2. Dodjeljivanje bodova za zadatak otvorenog tipa

U ovom dijelu se koristi podatkovni skup u kojem su za svakog studenta zabilježene ocjene koje su oni dali četvorici drugih studenata za njihove radove.

Za svakog studenta se nalaze četiri druga studenta koja su ocijenila njegov rad. Zatim se kao težine postavljaju koeficijenti determinacije izračunati u prvom dijelu. Bodove koji će biti dodijeljeni studentu za njegov rad računamo kao težinski prosjek bodova koje mu je određeni student dao pomnoženih s koeficijentom tog ocjenjivača. Odabran je ovakav pristup zato što koeficijent determinacije poprima veću vrijednost što je pojedini student točnije ocijenio kalibracijski test, odnosno maksimalnu vrijednost 1 ako je ocjenjivanje kalibracijskog testa bilo u potpunosti točno. To znači da ako je vrijednost koeficijenta determinacije nekog studenta jednaka 1, ocjena koju on dodijeli za tuđi rad ima puno veću težinu, odnosno referentnija je od ocjene koju dodijeli student s niskom točnosti ocjenjivanja. Ovakav pristup omogućuje puno precizniju i pravedniju evaluaciju studentskog rada, uzimajući u obzir varijacije u ocjenjivanju.

Kôd 9.2 prikazuje ključne dijelove ovog algoritma. Prvo se iterira se po podatkovnom skupu koji sadrži informacije o tome koliko je koji student dodijelio bodova kojem radu [1]. Zatim se u [2] izvlače svi radovi drugih studenata koje je student ocijenio te se u [3] izvlači jedinstveni identifikator studenta ocjenjivača. U [4] se iterira po svim radovima koje taj student on ocijenio te se za svaki rad izdvaja jedinstveni identifikator studenta koji je autor tog rada [5]. U liniji [10] se izvlači ocjena koju je student čiji zapis gledamo dodijelio studentu čiji je rad u pitanju. Zatim se u liniji [12] dohvaća podatak o težini studenta koji je dodijelio ocjenu. U liniji [16] se sprema dodijeljena ocjena i težina kao uređeni par te ih se pohranjuje za studentski rad koji razmatramo. Za pohranu se koristi podatkovna struktura rječnik. Nakon što je ta petlja izvršena, u liniji [18] se iterira po svim studentskim radovima te se u liniji [19] računa konačan broj bodova kao težinski prosjek ocjena koje je taj rad dobio od drugih studenata.

1	`for _, row in peer_review_data.iterrows():`
2	`  jobs = [job for job in row["jobs"].replace("{", "").replace("}", "").split(",") if int(job) > 0]`
3	`  grader = row["id_student"]`
4	`  for idx, job in enumerate(jobs):`
5	`  student = jobs_students[jobs_students["id_job"] == int(job)]["id_student"].values[0]`

```
6
7                if student not in dict_students:
8                        dict_students[student] = []
9
10               grade = row.iloc[idx - 4]
11               try:
12                       weight = calibration_data[calibration_data["id_student"]
                                    == grader]["weights"].values[0]
13               except IndexError:
14                       break
15
16               dict_students[student].append((grade, weight))
17  students_calculated_grades = {}
18  for student, grades in dict_students.items():
19          students_calculated_grades[student] = sum([grade * weight for grade,
                         weight in grades]) / sum( [weight for _, weight in grades])
```

Kôd 9.2 - Algoritam za dodjeljivanje bodova za zadatak otvorenog tipa

U ovom dijelu se za koeficijente ne koristi linearnu transformaciju već zaglađivanje koje se radi na sljedeći način:

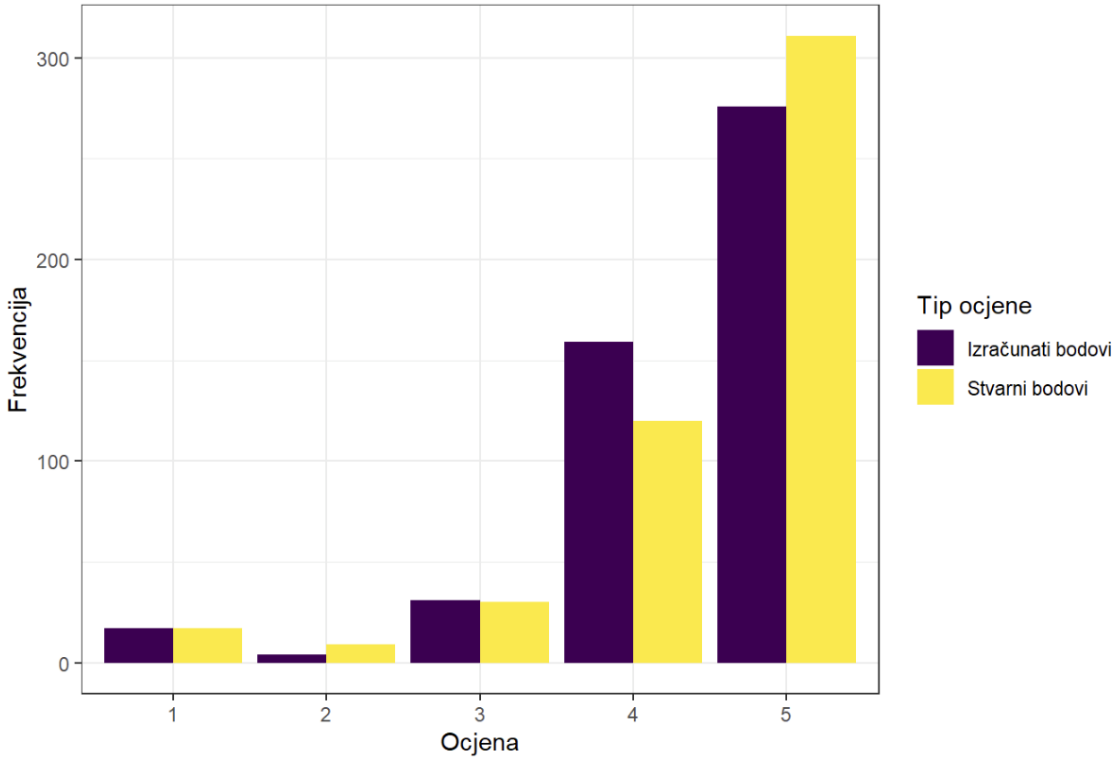
```
scaled_weights = [0.5 * (weight + mean_weight) for weight in weights]
```

Svaki koeficijent se zbroji s aritmetičkom sredinom težina, a zatim se taj zbroj pomnoži s 0.5. Time se pokušava postići isti efekt kao i u prvom dijelu, a to je ublažavanje koeficijenata.

Kada se usporede bodovi koje bi svaki student trebao dobiti primjenom ovog algoritma s onima koje je stvarno dobio na toj laboratorijskoj vježbi, podudaranje je 80%.

Na *Slika 9.2* vidi se graf usporedbe frekvencije ocjena koje su studenti dobili primjenom predloženog algoritma s njihovim stvarnim ocjenama. Jedina veća razlika je to što je predloženim algoritmom nešto više studenata dobilo ocjenu četiri, a nešto manje studenata ocjenu pet. Analizom podataka vidi se da su to upravo studenti na samoj granici između četiri i pet bodova, što sugerira da je algoritam nastavnika nešto blaži pri zaokruživanja ocjena.

Podatci iz *Tablica 9.1* prikazuju sumu bodova osvojenih predloženim algoritmom sa sumom stvarnih osvojenih bodova. Vidi se da se sume razlikuju samo za dva.



Slika 9.2 - Prikaz frekvencije bodova dodijeljenih predloženim algoritmom i stvarnih bodova

Suma osvojenih bodova algoritmom predloženim u ovom radu	Suma osvojenih bodova temeljem algoritma korištenog na predmetu
2143	**2145**

Tablica 9.1 - Usporedba sume osvojenih bodova studenata

Zaključak

U ovom radu istražene su mogućnosti korištenja velikih jezičnih modela (LLM) u ocjenjivanju zadataka otvorenog tipa, s posebnim fokusom na zadatke koji uključuju oblikovanje ER i relacijskih modela podataka temeljenih na opisima poslovnih procesa. Cilj je bio procijeniti kvalitetu i relevantnost ocjena i povratnih informacija koje pružaju LLM-ovi te usporediti te rezultate s ocjenama i komentarima dobivenim od strane studenata i nastavnika.

Reprezentativan broj studentskih rješenja ocijenjen je od strane dva velika jezična modela, ChatGPT-a i Bard-a. Nakon toga je provedena detaljna analiza i usporedba rezultata s ocjenama studenata-vršnjaka i nastavnika. Dobiveni rezultati pokazali su da LLM-ovi imaju sposobnost prepoznavanja ključnih elemenata i struktura u studentskim rješenjima, no ta sposobnost još uvijek nije na razini ljudskih ocjenjivača. LLM-ovi su pokazali tendenciju davanja viših ocjena studentima, često ne prepoznajući greške i ne pružajući kvalitetne komentare koji bi potkrijepili njihove odluke pri dodjeli ocjena.

U sklopu rada predložena je mjera za ocjenjivanje sposobnosti LLM-ova u ocjenjivanju zadataka otvorenog tipa, koristeći koeficijent determinacije i srednju kvadratnu grešku. Također je implementiran algoritam za pravedno dodjeljivanje bodova na temelju točnosti ocjenjivanja kalibracijskog testa.

Zaključno, iako LLM-ovi pokazuju značajan potencijal u ocjenjivanju zadataka otvorenog tipa, ne mogu biti zamjena za ljudske ocjenjivače. Optimalan pristup bila bi kombinacija ljudske prosudbe i automatiziranih alata, što bi osiguralo efikasnost, ali i točnost i kvalitetu ocjena i povratnih informacija.

Literatura

[1] ChatGPT (2024.) Pristupljeno 15.3 iz

https://chatgpt.com/

[2] Bard/Gemini (2024.) Pristupljeno 15.3. iz

https://gemini.google.com/app

[3] Kasneci, Enkelejda, et al. "ChatGPT for good? On opportunities and challenges of large language models for education." Learning and individual differences 103 (2023): 102274

[4] Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A. and Gama, K., 2023, September. Large language models for education: Grading open-ended questions using chatgpt. In Proceedings of the XXXVII Brazilian Symposium on Software Engineering (pp. 293-302).

[5] Pristupljeno 30.5.2024. iz

https://www.researchgate.net/figure/Visualization-of-SSE-SSR-SST_fig17_322398615

[6] Rstudio IDE (2024.) Pristupljeno 10.3.2024.

https://www.rstudio.com/categories/rstudio-ide/

[7] R (2024.) Pristupljeno 10.3.2024. iz

https://cran.r-project.org/bin/windows/base/

[8] Ggplot2 (2024.) Pristupljeno 10.3.2024 iz

https://ggplot2.tidyverse.org/

[9] PyCharm (2024.) Pristupljeno 1.4.2024. iz

https://www.jetbrains.com/pycharm/

[10] Python (2024.) Pristupljeno 1.4.2024. iz

https://www.python.org/

[11] Pandas (2024.) Pristupljeno 1.4.2024 iz

https://pandas.pydata.org/

[12] NumPy (2024.) Pristupljeno 1.4.2024. iz

https://pypi.org/project/numpy/

[13] Microsoft Excel (2024.) Pristupljeno 10.3.2024 iz

https://www.python.org/

Sažetak

Naziv: Korištenje velikih jezičnih modela u ocjenjivanju zadataka otvorenog tipa

U ovom radu istražuje se primjena velikih jezičnih modela u ocjenjivanju zadataka otvorenog tipa. Konkretno, korišteni su modeli ChatGPT i Bard. Rad opisuje karakteristike velikih jezičnih modela te statističke mjere korištene za analizu, uključujući sumu kvadrata i koeficijent determinacije. Nadalje, prikazan je podatkovni skup korišten u istraživanju, s postupkom prikupljanja relevantnih podataka. Opisane su korištene tehnologije i alati koji su omogućili provedbu analize. Rad također nudi interpretaciju rezultata dobivenih analizom ocjenjivanja i povratnih informacija generiranih od strane velikih jezičnih modela. Uz to, predložena je mjera za ocjenjivanje sposobnosti vrednovanja zadataka ovog tipa, kao i implementacija algoritma za dodjeljivanje bodova temeljenih na kalibracijskom testu.

Ključne riječi:

- Veliki jezični modeli
- ChatGPT
- Bard
- Zadatci otvorenog tipa
- Suma kvadrata
- Koeficijent determinacije

Summary

Title: Using Large Language Models in evaluating open-ended assignments

The purpose of this work is to determine the quality of large language models in evaluating open-ended assignments. More specifically, we are using ChatGPT and Bard. This paper describes characteristics of large language models in detail, along with various statistical measures used for analysis, such as sum of squares or coefficient of determination. Furthermore, the dataset used for the analysis is presented along with the process of acquiring and extracting all the relevant data from it. Technologies and other tools used analysis implementation are also described. The paper provides interpretation of the results of the analysis of the given dataset acquired through the grades and feedback generated by large language models. The measure for the evaluation the ability of grading tasks of this structure is proposed, along with the implementation of the algorithm for the distribution of points based on calibration test.

Keywords:

- Large language models
- ChatGPT
- Bard
- Open-ended assignments,
- Sum of squares
- Coefficient of determination