

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET



Svetozar Iković

**PRIMENA VEŠTAČKE INTELIGENCIJE U
OBRAZOVANJU: VELIKI JEZIČKI MODELI
ZA TESTIRANJE I OCENJIVANJE
STUDENATA**

master rad

Beograd, 2026.

Mentor:

prof. dr Jelena GRAOVAC

Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

prof. dr Milena VUJOŠEVIĆ JANIČIĆ

Univerzitet u Beogradu, Matematički fakultet

prof. dr Saša MALKOV

Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

Zahvalnica

Zahvaljujem se mentorki, prof. dr. Jeleni Graovac, na pomoći i podršci tokom celokupne izrade ovog rada, od početne ideje do realizacije, na usmeravanju istraživanja i na strpljivom čitanju svake verzije teksta teze.

Naslov master rada: Primena veštačke inteligencije u obrazovanju: veliki jezički modeli za testiranje i ocenjivanje studenata

Rezime: Ubrzani razvoj veštačke inteligencije i velikih jezičkih modela (engl. *large language models*, LLM) menja praksu provere znanja u visokom obrazovanju, naročito u ocenjivanju pitanja otvorenog tipa. Tradicionalno ocenjivanje je vremenski zahtevno, sklono nedoslednosti i teško primenljivo na velike grupe studenata. Veliki jezički modeli mogu da analiziraju i tumače značenje slobodnog teksta, pa povratna informacija može biti brža, doslednija i češća. U ovom radu predstavljen je EduGrader, platforma za ocenjivanje odgovora otvorenog tipa koju je autor razvio i objavio kao softver otvorenog koda. Platforma objedinjuje pet velikih jezičkih modela: GPT-4o, GPT-4o-mini, DeepSeek-Chat, GPT-5 i DeepSeek-Reasoner. EduGrader podržava tri konfiguracije strogosti ocenjivanja (blagu, neutralnu, strogu) i uz numeričku ocenu daje sažeto obrazloženje: šta je u odgovoru ispravno, šta nedostaje i gde je greška. Sistem radi u dva režima: režimu sa zadatim referentnim odgovorom (engl. *provided-reference-answer*, PRA), vođenom referentnim odgovorima nastavnika, i režimu sa generisanim referentnim odgovorom (engl. *generated-reference-answer*, GRA), koji referentne odgovore automatski generiše iz nastavnih materijala. Svi eksperimenti sprovedeni su na srpskom jeziku, koji spada u jezike sa ograničenim resursima i znatno je manje zastupljen u podacima za obučavanje savremenih modela, čime je ispitivanje prošireno izvan dominantnog engleskog jezičkog konteksta. EduGrader je evaluiran na 583 autentična studentska odgovora sa pet univerzitetskih kurseva na dve institucije, obuhvatajući i tekstualna obrazloženja i programski kod. U najboljoj konfiguraciji sistem postiže Pirsonov koeficijent korelacije od 0,89 sa ljudskim ocenjivačima. Zamišljen je kao alat za podršku ocenjivanju: preuzima rutinski deo posla, a ne zamenjuje nastavnika.

Ključne reči: veliki jezički modeli, automatsko ocenjivanje, pitanja otvorenog tipa, referentni odgovor, srpski jezik, jezik sa ograničenim resursima, procena znanja u visokom obrazovanju, EduGrader

Sadržaj

1	Uvod	1
1.1	Motivacija	2
1.2	Sistem EduGrader i doprinosi rada	2
2	Pregled oblasti i srodni radovi	6
2.1	Veštačka inteligencija u obrazovanju	6
2.2	Empirijske studije ocenjivanja jezičkim modelima	7
2.3	Pozicioniranje ovog rada	9
3	Veb aplikacija EduGrader	11
3.1	Klijentska aplikacija	12
3.2	Serverska aplikacija	15
3.3	Integracija jezičkih modela	15
4	Sistem za automatizovanu evaluaciju modela	17
4.1	Format ulaznih podataka	17
4.2	Faze obrade	18
4.3	Automatizacija eksperimenata i analiza rezultata	19
4.4	Reprodukcija eksperimenata	20
5	Metodologija evaluacije	24
5.1	Prikupljanje podataka	24
5.2	Etička pitanja i zaštita podataka	26
5.3	Metrike i evaluacija	27
5.4	Modeli i konfiguracije	29
5.5	Ograničenje GRA režima za modele DeepSeek	30
5.6	Treći režim: ocenjivanje na osnovu nastavnog materijala	30

6	Rezultati	32
6.1	Performanse modela	32
6.2	Poređenje PRA i GRA režima	34
6.3	Korelacija sa ljudskim ocenama po skupovima podataka	35
6.4	Dodatni eksperimenti	38
7	Diskusija	40
7.1	Poređenje sa prethodnim istraživanjima	40
7.2	Ograničenja	41
7.3	Pravci budućeg rada	43
8	Zaključak	45
A	Šabloni promptova i primeri ocenjivanja	47
B	Upotreba velikih jezičkih modela u izradi rada	52
	Literatura	54

Glava 1

Uvod

Ocenjivanje retko predstavlja najzanimljiviji deo nastave, ali oblikuje način na koji studenti uče. Pravovremena i konkretna povratna informacija omogućava studentu da uoči i ispravi greške u razumevanju pre nego što se one ustale, dok zakašnela ili nedovoljno jasna povratna informacija umanjuje značaj provere znanja i svodi je na formalnost. Na velikim kursevima nastavnici i saradnici ocenjuju stotine odgovora otvorenog tipa u kratkim rokovima. Kratka objašnjenja i mini-eseji, za razliku od pitanja sa ponuđenim odgovorima, ne mogu se proveriti prostim uparivanjem obrazaca. Oni zahtevaju tumačenje: prepoznavanje delimično ispravnog rezonovanja, razlikovanje jezičkih nepreciznosti od konceptualnih praznina i odluku o tome koliko neki propust treba da utiče na ocenu. Takvo ocenjivanje je skupo, teško primenljivo na velike grupe studenata i sklono nedoslednosti: čak i uz detaljna uputstva, dva ocenjivača mogu isti odgovor protumačiti različito, naročito kada je odgovor dvosmislen, pisan u žurbi ili strukturno nejasan [4, 12, 27]. Nastavnika to opterećenje odvraća od pitanja zasnovanih na objašnjenjima, a student povratnu informaciju dobija prekasno ili je ne dobija uopšte.

Raniji sistemi za automatsku proveru znanja radili su samo kada su odgovori bili strogo ograničeni: testovi sa ponuđenim odgovorima, pitanja dopunjavanja ili kratki odgovori sa predvidljivom formulacijom [3, 23, 26]. Napredak u obradi prirodnog jezika proširio je te mogućnosti, ali su se mnogi pristupi i dalje oslanjali na ručno projektovanje atributa, uske domene ili velike označene skupove podataka [2, 5, 24]. Savremeni veliki jezički modeli (engl. *large language models*, LLM) tumače odgovore u slobodnom tekstu. Time je postalo moguće graditi alate koji ocenjuju odgovore otvorenog tipa, obrazlažu ocenu prirodnim jezikom i ukazuju na ono što u odgovoru nedostaje, a ne samo na ono što je pogrešno [14, 34].

1.1 Motivacija

Cilj ovog rada je ispitivanje mogućnosti pouzdane primene velikih jezičkih modela kao podrške u ocenjivanju. Ručno ocenjivanje troši veliki deo nastavničkog vremena, pa povratna informacija kasni, a provere znanja se organizuju ređe nego što bi bilo korisno. Automatizovan sistem za ocenjivanje menja tu ravnotežu. Provere znanja mogu biti češće, pa student kontinuirano prati napredak i prepoznaje oblasti koje zahtevaju dodatni rad. Analiza odgovora stiže dok je gradivo još sveže. Vreme koje bi otišlo na rutinski deo ocenjivanja nastavnik usmerava na individualnu podršku studentima.

Upotreba velikih jezičkih modela u ocenjivanju istovremeno otvara praktična i pedagoška pitanja. Ovaj rad razmatra četiri:

1. Koliko su ocene velikih jezičkih modela usklađene sa nastavničkim i koliko su dosledne kroz različite predmete i tipove pitanja?
2. Mogu li nastavnici kalibrisati strogost ocenjivanja prema očekivanjima konkretnog kursa?
3. Šta se dešava kada potpun referentni odgovor nastavnika nije dostupan?
4. Kako sistem za automatizovano ocenjivanje projektovati tako da nastavnik zadrži kontrolu nad konačnom ocenom?

Ova pitanja se obrađuju kroz ceo rad, a odgovori na njih sažeti su u zaključku (poglavlje 8).

1.2 Sistem EduGrader i doprinosi rada

U ovom radu predstavljen je EduGrader, platforma za ocenjivanje pitanja otvorenog tipa u visokom obrazovanju, pod nadzorom nastavnika, koju je autor projektovao i razvio u okviru ovog rada [10, 11, 9]. EduGrader objedinjuje više velikih jezičkih modela, GPT-4o,¹ GPT-4o-mini,² DeepSeek-Chat,³ i dva modela za rezonovanje (engl. *reasoning models*), GPT-5⁴ i DeepSeek-Reasoner,⁵ u jedinstven tok ocenjivanja u kome poslednju reč ima nastavnik.

¹<https://platform.openai.com/docs/models/gpt-4o>

²<https://platform.openai.com/docs/models/gpt-4o-mini>

³<https://api-docs.deepseek.com/>

⁴<https://platform.openai.com/docs/models/gpt-5>

⁵https://api-docs.deepseek.com/guides/reasoning_model

Sistem podržava tri konfiguracije strogosti ocenjivanja: blagu, neutralnu i strogu. Podela odgovara nastavnoj praksi, u kojoj neki kursevi nagrađuju delimično ispravno rezonovanje, dok drugi zahtevaju preciznu terminologiju i potpuna rešenja.

EduGrader nudi dva komplementarna režima ocenjivanja:

- **Režim sa zadatim referentnim odgovorom** (engl. *provided-reference-answer*, PRA), u kome se studentski odgovori porede sa referentnim odgovorima koje je pripremio nastavnik.
- **Režim sa generisanim referentnim odgovorom** (engl. *generated-reference-answer*, GRA), u kome sistem najpre generiše referentni odgovor na osnovu nastavnih materijala, beležaka sa predavanja, skripti i udžbenika, a zatim studentske odgovore ocenjuje u odnosu na tako generisani standard.

Izlaz u oba režima je numerička ocena praćena obrazloženjem koje ukazuje na jake strane odgovora, koncepte koji nedostaju i praznine u rezonovanju. Takva povratna informacija studentu je upotrebljiva, a nastavniku proverljiva, za razliku od čiste numeričke ocene bez obrazloženja.

Model ocenu izražava na skali od 0 do 100, jer finija podela skale olakšava razlikovanje bliskih nivoa kvaliteta odgovora. Dobijena vrednost se zatim preslikava na skalu od 0 do 10, na kojoj nastavnici dodeljuju poene za svaki pojedinačni odgovor. Reč je, dakle, o poenima po pitanju, a ne o zaključnoj oceni na kursu (od 5 do 10). Sve ocene u ovom radu, i systemske i nastavničke, prikazane su na skali 0–10, osim tamo gde je naznačeno drugačije.

Svi eksperimenti sprovedeni su na srpskom jeziku, koji spada u jezike sa ograničenim resursima i znatno je manje zastupljen u podacima za obučavanje savremenih modela, čime je ispitivanje prošireno izvan dominantnog engleskog jezičkog konteksta. Sistem je evaluiran na 583 autentična studentska odgovora prikupljena na dva univerziteta i pet kurseva, od kratkih činjeničkih odgovora i otvorenih tekstualnih objašnjenja do odgovora u obliku koda. Rezultati pokazuju da PRA režim daje stabilnije ishode od GRA režima i da modeli za rezonovanje daju ocene najbliže nastavničkim. U najboljoj konfiguraciji (GPT-5, blaga strogost) EduGrader postiže Pirsonov koeficijent korelacije (engl. *Pearson correlation coefficient*, PCC) od 0,89 u odnosu na ljudsko ocenjivanje, a u neutralnoj konfiguraciji 0,88.

EduGrader je zamišljen kao alat za podršku ocenjivanju, a ne kao zamena za nastavnika. Namenjen je smanjenju ponavljajućeg dela ocenjivanja, ali ljudski

nadzor ostaje neophodan za dvosmislena pitanja, očekivanja specifična za disciplinu i granične slučajeve. Nastavnik pri tome i dalje čita svaki odgovor, pa pomoć nije u preskakanju čitanja, već u tri stvari. Obrazloženje koje sistem daje uz ocenu je nacrt povratne informacije koju bi nastavnik studentu ionako pisao. Predložena ocena je druga, nezavisna procena, koja pomaže doslednosti kroz stotine odgovora u istom roku. Konačno, kod odgovora gde se nastavnik sa sistemom slaže dovoljna je potvrda, pa se vreme ulaže u granične i sporne slučajeve. Koliko vremena se time stvarno štedi u ovom radu nije mereno. Zato je sistem projektovan tako da je čovek njegov sastavni deo (engl. *human-in-the-loop*) i objavljen kao softver otvorenog koda [9], kako bi bio proverljiv i prilagodljiv drugim institucijama.

Doprinosi rada su sledeći:

1. **Veb aplikacija EduGrader**, koju je autor projektovao i implementirao: nastavnički i studentski modul, serverski modul za ocenjivanje koji čuva prompt, odgovor modela i izdvojenu ocenu za svaki ocenjeni odgovor, i integracioni sloj koji više provajdera (engl. *provider*) jezičkih modela pokriva jednim interfejsom.
2. **Sistem za automatizovanu evaluaciju modela**, koji je autor razvio: ponovljiv postupak (engl. *pipeline*) koji nad ručno pripremljenim ispitnim podacima pokreće ocenjivanje za proizvoljnu kombinaciju modela, konfiguracije strogosti i režima i računa metrike korelacije i slaganja sa ocenama nastavnika.
3. **Skup podataka na srpskom jeziku**: 583 autentična, pseudonimizovana studentska odgovora sa pet kurseva na dve institucije, sa nastavničkim ocenama i referentnim odgovorima. Studentski odgovori nisu javno objavljeni, već su dostupni na zahtev.
4. **Sistematska evaluacija**: pet modela u tri konfiguracije strogosti i dva režima ocenjivanja, dopunjena eksperimentima sa nastavnim materijalom u promptu i sa temperaturom, parametrom koji upravlja slučajnošću pri uzorkovanju odgovora modela.
5. **Empirijski nalazi**: referentni odgovor nastavnika dosledno poboljšava i korelaciju i slaganje sa ljudskim ocenjivačima, modeli za rezonovanje su im najbliži, a stroga konfiguracija po pravilu pogoršava oba.

Rezultati istraživanja predstavljeni su na konferenciji *Artificial Intelligence Srpske akademije nauka i umetnosti (SANU)* (Beograd, oktobar 2025) [10] i prihvaćeni za objavljivanje u časopisu *Computer Science and Information Systems (ComSIS)* [11]. Ovaj rad je proširena verzija tog članka: detaljnije opisuje aplikaciju i evaluacionu infrastrukturu, a rezultate dopunjuje analizom po kursevima i eksperimentima sa nastavnim materijalom u promptu i sa temperaturom.

Ostatak rada organizovan je na sledeći način. Poglavlje 2 daje pregled srodnih istraživanja primene veštačke inteligencije u ocenjivanju i pozicionira ovaj rad u odnosu na njih. Poglavlje 3 opisuje veb aplikaciju EduGrader i sloj za integraciju jezičkih modela, a poglavlje 4 sistem za automatizovanu evaluaciju modela kojim su sprovedeni eksperimenti. Poglavlje 5 predstavlja metodologiju evaluacije: skupove podataka, zaštitu podataka, metrike i konfiguracije modela. Poglavlje 6 donosi rezultate, uključujući poređenje PRA i GRA režima kroz modele i konfiguracije strogosti, uticaj konfiguracije strogosti i korelacije po pojedinačnim kursevima. U poglavlju 7 rezultati su upoređeni sa prethodnim istraživanjima i razmotreni su pravci budućeg rada, dok poglavlje 8 zaključuje rad. Dodatak A sadrži šablone promptova i reprezentativne primere ocenjivanja. Dodatak B navodi način upotrebe velikih jezičkih modela pri izradi rada.

Glava 2

Pregled oblasti i srodni radovi

Pregled polazi od šireg konteksta primene veštačke inteligencije u obrazovanju i pitanja transparentnosti i odgovornosti koja automatsko ocenjivanje otvara. Zatim su prikazane empirijske studije ocenjivanja jezičkim modelima, od eseja i kratkih odgovora do programskog koda, sa naglaskom na to šta je u svakoj studiji model dobijao uz pitanje i kojim metrikama je meren. Na kraju su te studije sažete po dimenzijama eksperimenta i naznačeno je šta ovaj rad u odnosu na njih dodaje.

2.1 Veštačka inteligencija u obrazovanju

Primena veštačke inteligencije u obrazovanju razvijala se od rane automatizacije zasnovane na pravilima ka adaptivnim sistemima koji tumače i generišu složene odgovore nalik ljudskim. Pod veštačkom inteligencijom se podrazumeva ju računarski sistemi projektovani da obavljaju kognitivne funkcije poput učenja, rezonovanja i rešavanja problema [1], a u obrazovnom kontekstu skup tehnologija koje omogućavaju personalizovanija i adaptivnija okruženja za učenje [33]. Sistematski pregledi im pripisuju potencijal da unaprede ishode učenja, prošire dostupnost obrazovanja i efikasnije iskoriste nastavne resurse [32, 33]. Veliki jezički modeli tu donose promenu u odnosu na ranije pristupe zasnovane na obradi prirodnog jezika: ne zahtevaju ručno projektovane attribute ni obučavanje za konkretan domen, već obrađuju slobodan tekst i tumače njegovo značenje na višem nivou apstrakcije. Zato se ispituju za nastavu usmerenu na učenika, automatsko generisanje sadržaja i nove oblike provere znanja [14].

Izvan numeričkog ocenjivanja, sistemi veštačke inteligencije generišu adaptivne aktivnosti učenja, daju personalizovanu povratnu informaciju i podržavaju forma-

tivno i sumativno ocenjivanje prilagođeno karakteristikama učenika [7, 17]. Intelligentni tutorski sistemi (engl. *intelligent tutoring systems*) nastoje da aproksimiraju individualnu nastavnu podršku uz veću doslednost i skalabilnost evaluacije [6]. Kritičke perspektive upozoravaju na „zamagljivanje” odgovornosti u automatskom ocenjivanju, gde odluke postaju netransparentne, a odgovornost razvodnjena [28]. Transparentnost, objašnjivost, privatnost podataka i kvalitet povratnih informacija zato ostaju centralne teme ocenjivanja uz veštačku inteligenciju [16].

2.2 Empirijske studije ocenjivanja jezičkim modelima

Empirijske studije ocenjivanja uz podršku velikih jezičkih modela daju mešovite rezultate. Kuli i Jusuf [15] ispitali su ocenjivanje uz pomoć ChatGPT-a na 25 studentskih radova iz generacije od približno 100 studenata druge godine na kursu iz društvenih nauka, gde su radove nezavisno ocenili i ljudski ocenjivači i model. Dobijena je umerena korelacija između ocena veštačke inteligencije i ljudskih ocena (koeficijent korelacije približno 0,45): model aproksimira opšte obrasce ocenjivanja, ali pojedinačne evaluacije se i dalje razlikuju. Autori ističu da ChatGPT ubrzava ocenjivanje i daje korisnu povratnu informaciju, ali da mu izostaje kontekstualno razumevanje nijansiranih studentskih argumenata. Upozoravaju i na rizik produbljivanja obrazovnih nejednakosti i na potrebu za institucionalnim okvirima za transparentnost, nadzor i ublažavanje pristrasnosti pri uvođenju takvih alata. Za ovaj rad ta studija daje donju tačku poređenja: ona meri eseje iz društvenih nauka, dok se ovde ocenjuju kratki odgovori iz STEM oblasti (nauka, tehnologija, inženjerstvo i matematika), za koje literatura očekuje veću korelaciju.

Lundgren [18] je ispitao GPT-4 na 60 eseja master nivoa iz političkih nauka koje su prethodno ocenili univerzitetski nastavnici. Prosečne ocene modela bile su usklađene sa nastavničkim, ali je pouzdanost među ocenjivačima (između modela i ljudi) bila niska: model prati agregirane trendove, a ne fino ljudsko rasuđivanje. Model je pri ocenjivanju izbegavao rizik i oslanjao se na opšte karakteristike eseja, poput kvaliteta jezika, umesto na detaljne rubrike (engl. *rubric*). Izmene prompta i uputstava za ocenjivanje donele su samo ograničena poboljšanja, pa su tadašnji sistemi bili slabo osetljivi na suptilne varijacije u kriterijumima evaluacije. Upravo taj nalaz motiviše eksperiment sa tri konfiguracije strogosti u ovom radu: umesto

da se pretpostavi da instrukcija o strogosti deluje, meri se da li i koliko ona pomera ocene i njihovu korelaciju sa nastavničkim.

Jukjevič [13] je ChatGPT primenio na ocenjivanje devet programerskih zadataka na petnaestonedeljnom kursu programiranja u *Python*-u sa 67 studenata osnovnih studija kognitivnih nauka, gde su radove nezavisno ocenili i nastavnik i model. Korelacija sa nastavničkim ocenama bila je snažno pozitivna, mada je model dodeljivao nešto strože ocene. Ponovljivost kroz višestruka pokretanja bila je visoka, sa zanemarljivim varijacijama između iteracija. Uz numeričku ocenu, ChatGPT je procenjivao ispravnost koda, proveravao poštovanje smernica PEP 8 (zvaničnog stilskeg vodiča za *Python* kod) i za nekoliko sekundi generisao detaljnu povratnu informaciju. Autor navodi i ograničenja: halucinirane greške, troškove korišćenja API-ja (engl. *application programming interface*) i povremene nedoslednosti u ocenjivanju koje traže proveru nastavnika. Zaključak je da je ChatGPT koristan dopunski alat za procenu koda i generisanje povratnih informacija, ali da ljudski nadzor ostaje neophodan za pouzdanost i pravednost. Dva nalaza te studije ovde se proveravaju na drugom jeziku i novijim modelima: sistematsko odstupanje od nastavnika prati se kroz odnos prosečnih ocena (engl. *mean ratio*), a ponovljivost kroz eksperiment sa temperaturom.

Olivos i dr. [21] ispitali su saradnju nastavnika i modela na 70 rukom pisanih ispita sa uvodnog kursa sociologije, gde su GPT-4 i GPT-4o-mini korišćeni za transkripciju rukopisa i ocenjivanje pitanja esejskog tipa. Ocene modela su bile snažno korelisane sa ocenama ljudskog ocenjivača, naročito kada je modelu uz pitanje dat i šablonski odgovor (engl. *template answer*). Pošto su odstupanja ostala, autori modelu namenjuju ulogu „drugog ocenjivača” koji označava nedoslednosti za proveru, a ne zamene za ljudsku procenu. Razlika između ocenjivanja sa referentnim odgovorom i bez njega tamo je zabeležena usput; ovde je ona eksperimentalna promenljiva: PRA i GRA režim pokrenuti su nad istim podacima i modelima.

Vecler i dr. [31] uporedili su osam nastavnika sa dve verzije ChatGPT-a (3.5 i 4o) na 186 eseja sa uvodnog kursa psihologije, ocenjenih po istoj detaljnoj rubrici, koristeći Bland-Altmanovu i regresionu analizu. ChatGPT 3.5 je u proseku ocenjivao više od nastavnika, dok su prosci verzije 4o bili bliži ljudskim, ali su obe verzije bile blaže od nastavnika kod slabijih radova i strože kod boljih, što je proporcionalno sistematsko odstupanje. Autori zaključuju da model i nastavnici istu rubriku primenjuju različito. Taj rezultat je razlog što se u ovom radu korelacija i slaganje posmatraju odvojeno: visoka korelacija može da postoji uz sistematsko

pomeranje ocena, pa se uz PCC prate i odnos prosečnih ocena i RMSE/STDDEV¹.

Ovi nalazi se uklapaju u širu sliku: pouzdanost ocenjivanja zavisi od tipa zadatka. Slaganje sa ljudskim ocenjivačima veće je kod strukturiranih odgovora (definicije, formule, fragmenti koda), a opada kod subjektivnih ili otvorenih pisanih zadataka [29, 30]. U radu [22] GPT-3 je na odgovorima otvorenog tipa profesionalaca iz industrije pokazao snažne korelacije sa ekspertskim ocenama kad su kriterijumi jasno definisani i kontekstualno utemeljeni.

2.3 Pozicioniranje ovog rada

Tabela 2.1 sažima empirijske studije iz prethodnog odeljka po dimenzijama koje su bitne za ovaj rad: vrsti zadatka, obimu podataka, modelima, ulozi referentnog odgovora, tretmanu strogosti i korišćenim metrikama. Podaci su preuzeti iz samih studija; gde studija neku dimenziju ne navodi, polje je označeno sa „n/p”.

Tabela 2.1: Srodne empirijske studije i ovaj rad po dimenzijama eksperimenta. Kolona „Ref. odgovor” označava šta je model dobijao uz pitanje pri ocenjivanju; kolona „Strogost” da li je strogost ocenjivanja bila kontrolisana promenljiva.

Studija	Zadatak i obim	Modeli	Ref. odgovor	Strogost	Metrike
Kuli i Jusuf [15]	eseji, društvene nauke, 25 radova	ChatGPT	n/p	ne	korelacija
Lundgren [18]	eseji, političke nauke, 60 eseja	GPT-4	rubrika	varijacije prompta	PCC, procenat slaganja, Koenova kapa
Jukjevič [13]	kod (<i>Python</i>), 9 zadataka, 67 studenata	ChatGPT	n/p	ne	korelacija, ponovljivost
Olivos i dr. [21]	eseji, sociologija, 70 ispita	GPT-4, GPT-4o-mini	sa šablonom i bez njega	ne	korelacija
Vecler i dr. [31]	eseji, psihologija, 186 eseja	GPT-3.5, GPT-4o	rubrika	ne	Bland-Altman, regresija
Ovaj rad	kratki odgovori i kod, STEM, srpski, 583 odgovora, 5 kurseva	5 (od toga 2 za rezonovanje)	PRA / GRA	3 konfiguracije	PCC, odnos prosečnih ocena, RMSE/STDDEV

Iz tabele se vide četiri praznine koje ovaj rad popunjava. Prva je jezik. Nijedna od navedenih studija ne ispituje ocenjivanje na jeziku sa ograničenim resursima

¹RMSE je koren srednje kvadratne greške (engl. *root mean square error*), a STDDEV standardna devijacija (engl. *standard deviation*).

niti jezik tretira kao okolnost eksperimenta; ovde su pitanja, studentski odgovori, referentni odgovori, nastavni materijali i promptovi na srpskom, pa rezultati govore o upotrebljivosti modela izvan engleskog konteksta u kome se obično ispituju. Druga je referentni odgovor kao promenljiva. Prethodne studije modelu ili daju rubriku ili je ne daju, a jedino Olivos i dr. beleže da šablonski odgovor popravljaja korelaciju; ovde su režimi sa referentnim odgovorom nastavnika (PRA) i sa generisanim referentnim odgovorom (GRA) pokrenuti nad istim odgovorima, modelima i konfiguracijama, pa se doprinos referentnog odgovora može izdvojiti od svega ostalog. Treća je strogost. Lundgren i Vecler i dr. uočavaju da model sistematski odstupa od nastavnika, ali odstupanje ne pokušavaju da kontrolišu; ovde su tri konfiguracije strogosti deo eksperimentalnog plana, a metrike razdvajaju korelaciju (PCC) od sistematskog odstupanja i greške (odnos prosečnih ocena, RMSE/STDDEV), pa se vidi da li instrukcija o strogosti menja redosled ocena, njihov nivo ili oboje. Četvrta su modeli: pregledane studije koriste ChatGPT, GPT-3.5, GPT-4 i GPT-4o, dok se ovde modeli opšte namene pod istim uslovima porede sa modelima za rezonovanje (GPT-5, DeepSeek-Reasoner). Numeričko poređenje postignutih vrednosti PCC-a sa onima iz literature dato je u odeljku 7.1.

Kroz svu navedenu literaturu provlači se isti zaključak: veliki jezički modeli su pomoćni alati koji automatizuju rutinske aspekte ocenjivanja, dok nastavnik ostaje nezamenljiv za pedagoško rasuđivanje, kontekstualno tumačenje i granične slučajeve [28, 33]. Preporukama iz literature najbliže odgovaraju pristupi koji spajaju automatizaciju sa nadzorom čoveka kao sastavnog dela sistema, podesivom strogošću ocenjivanja i razumljivom povratnom informacijom. Ta tri zahteva, kontrola nastavnika, transparentnost i prilagodljivost, oblikovala su i platformu EduGrader predstavljenu u ovom radu, a uloga „drugog ocenjivača” koju predlažu Olivos i dr. odgovara mestu koje model ima u njenom toku rada: model predlaže ocenu i obrazloženje, a konačnu odluku donosi nastavnik.

Glava 3

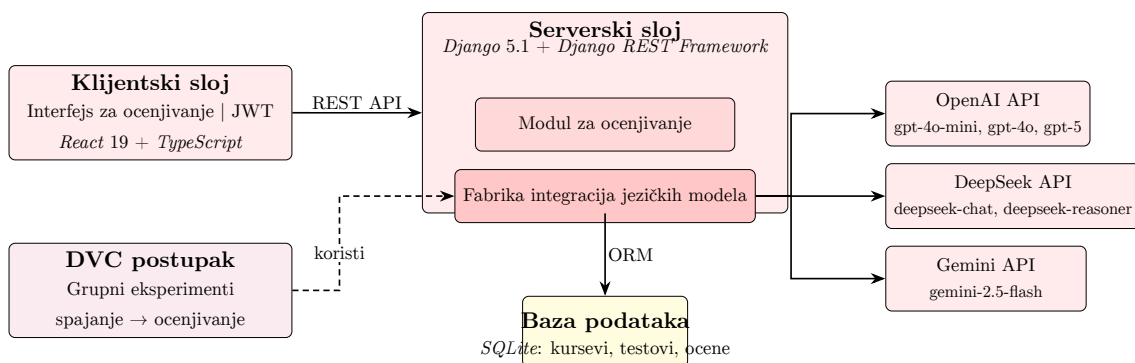
Veb aplikacija EduGrader

EduGrader je veb aplikacija za automatsko ocenjivanje odgovora otvorenog tipa sa studentskih ispita pomoću velikih jezičkih modela. To je deo sistema namenjen upotrebi u nastavi, organizovan u tri sloja (slika 3.1):

- **Klijentski sloj.** Jednostranična aplikacija (engl. *single-page application*, SPA) izgrađena u tehnologijama *React* 19 i *TypeScript*, koju poslužuje *Ngix* server. Prikazi za nastavnike i studente su odvojeni: nastavnici upravljaju kursevima, testovima i pitanjima i pregledaju rešene testove, a studenti rešavaju testove i prate svoje rezultate. Na prikazu rešenog testa, uz svaki odgovor stoje i ocena nastavnika i ocena koju je predložio jezički model; konačnu odluku donosi nastavnik. Autentifikacija se obavlja putem JWT-a (engl. *JSON Web Token*) sa kontrolom pristupa zasnovanom na ulogama.
- **Serverski sloj.** REST (engl. *representational state transfer*) API izgrađen u tehnologijama *Python* 3, *Django* 5.1 i *Django REST Framework*. Izlaže CRUD (engl. *create, read, update, delete*) operacije nad domenskim modelima i sadrži modul za ocenjivanje, odgovoran za sastavljanje promptova, pozivanje modela, izdvajanje ocene iz odgovora i trajno čuvanje rezultata.
- **Sloj baze podataka.** Realizovan kroz *Django* ORM (engl. *object-relational mapping*) nad bazom *SQLite*, sa entitetima za kurseve, testove, pitanja, studentske odgovore i ocene. *SQLite* je izabran kao razvojna i prototipska baza, dovoljna za lokalno pokretanje: umesto zasebnog serverskog procesa, biblioteka *SQLite* radi u samom procesu aplikacije, a celu bazu čuva u jednoj datoteci. *Django* ORM dopušta prelazak na serverski sistem za upravljanje bazom (npr. *PostgreSQL*) izmenom konfiguracije, bez izmena u kodu

aplikacije. Pored samih ocena čuvaju se i povratne informacije, promptovi i metapodaci, potrebni za kasnije analize.

Logika ocenjivanja i ponašanje modela razvijaju se nezavisno od korisničkog interfejsa; klijentski i serverski deo komuniciraju samo preko REST API-ja. Slika prikazuje i evaluacioni postupak u sistemu DVC (engl. *Data Version Control*)¹, opisan u poglavlju 4, koji koristi isti integracioni sloj.



Slika 3.1: Arhitektura sistema EduGrader. Klijentski sloj komunicira sa server-skim slojem preko REST API-ja. Serverski sloj pristupa jezičkim modelima kroz jedinstven integracioni sloj (projektni obrazac fabrika) koji apstrahuje API-je pojedinačnih provajdera. Sistem za automatizovanu evaluaciju koristi isti integracioni sloj za grupne eksperimente. Integracija Gemini nije korišćena u eksperimentima opisanim u ovom radu.

3.1 Klijentska aplikacija

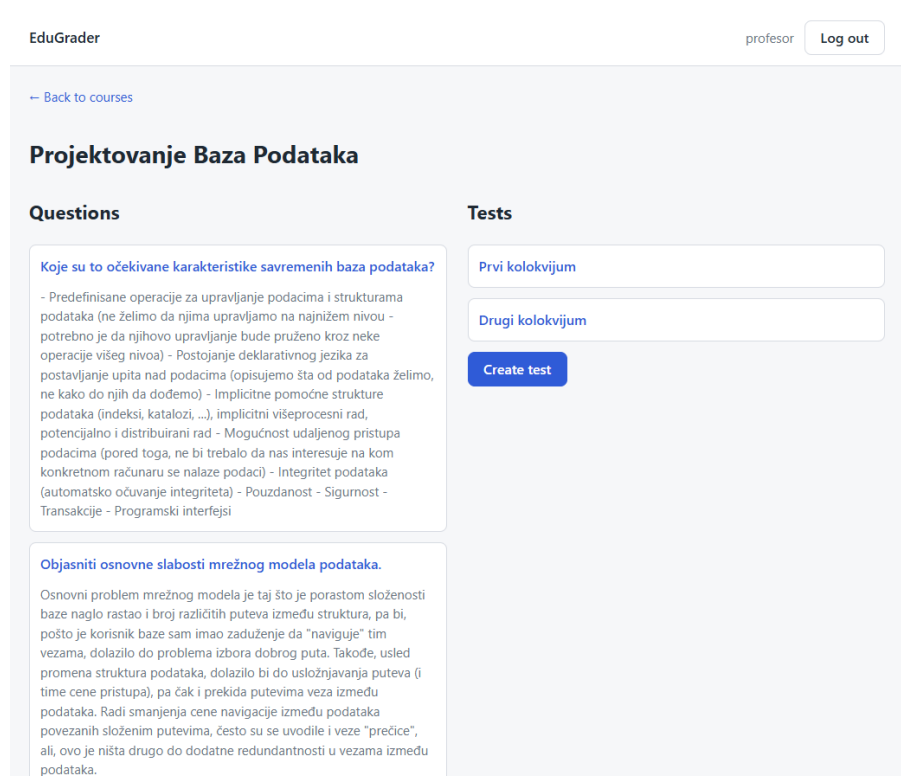
Klijentska aplikacija je jednostranična aplikacija koja koristi biblioteku *React* (verzija 19) i jezik *TypeScript*, uz *Vite* kao alat za izgradnju i razvojno okruženje. Rutiranje je rešeno bibliotekom *React Router*, kroz koju je aplikacija podeljena na dva modula sa odvojenim pravima pristupa:

- **Nastavnički modul** omogućava pregled predmeta, kreiranje testova i pitanja i pregled rešenih studentskih testova, sa prikazom nastavničke ocene uz predlog ocene jezičkog modela za svaki odgovor.
- **Studentski modul** omogućava pregled nadolazećih testova, izradu testa i uvid u rezultate već urađenih testova.

¹<https://dvc.org>

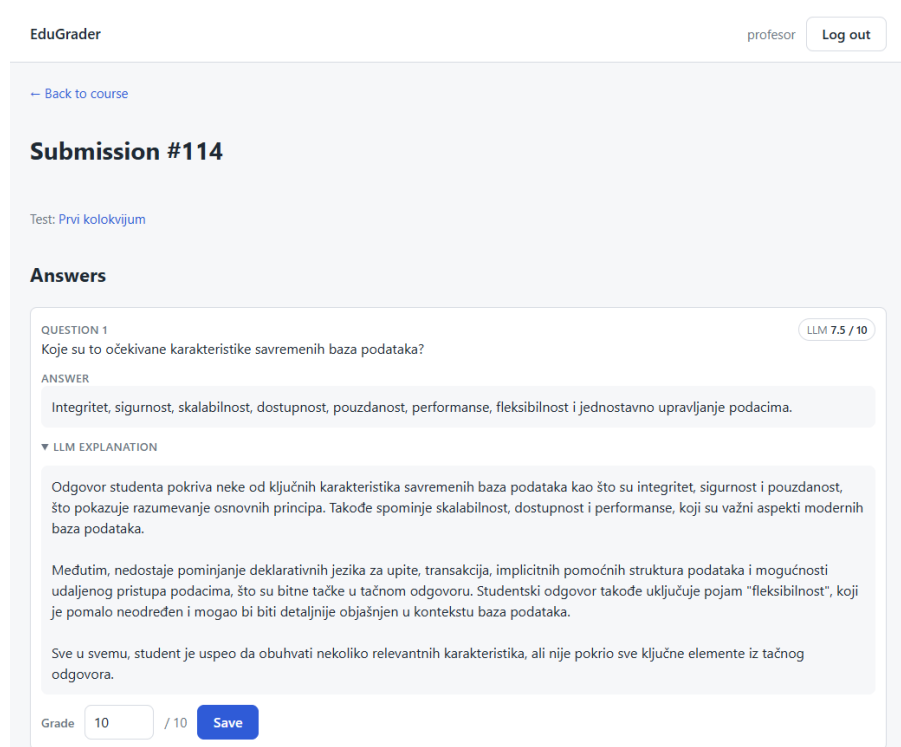
Slike 3.2–3.4 prikazuju tipičan tok rada. Nastavnik na stranici kursa unosi pitanja sa referentnim odgovorima i od njih sastavlja testove (slika 3.2). Kada student preda test (slika 3.4), nastavnik pri pregledu za svaki odgovor vidi predlog ocene jezičkog modela i, po potrebi, njegovo obrazloženje, a konačnu ocenu unosi sam (slika 3.3). Predlog modela je tako samo polazna tačka: nastavnik ga može prihvatiti ili prepraviti, u skladu sa principom iz uvoda da je čovek sastavni deo sistema.

Elementi interfejsa su na engleskom, dok su pitanja, referentni odgovori i studentski odgovori na srpskom, isto kao u svim eksperimentima u ovom radu.

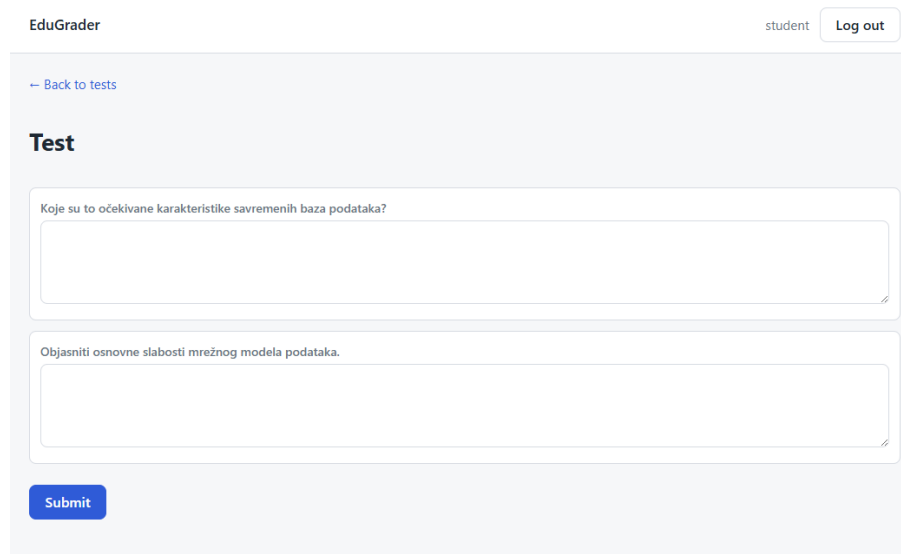


Slika 3.2: Nastavnički modul: stranica kursa sa pitanjima, njihovim referentnim odgovorima i listom testova.

Grupno ocenjivanje se u trenutnoj implementaciji ne pokreće iz korisničkog interfejsa: logika ocenjivanja izložena je kao funkcije serverskog modula za ocenjivanje, a eksperimenti iz ovog rada pokreću je kroz evaluacioni postupak u sistemu DVC, nad podacima pripremljenim u vidu CSV (engl. *comma-separated values*) datoteka. Izbor modela, režima ocenjivanja i konfiguracije strogosti zadaje se kroz parametarske datoteke tog postupka, a ne kroz veb interfejs.



Slika 3.3: Nastavnički modul: pregled rešenog testa. Uz svaki odgovor prikazan je predlog ocene jezičkog modela (ovde 7,5 od 10) sa rasklopljenim obrazloženjem, a konačnu ocenu unosi nastavnik u polje *Grade*. Na snimku je nastavnik uneo konačnu ocenu 10, različitu od predloga modela, što ilustruje da odluka ostaje na njemu.



Slika 3.4: Studentski modul: izrada testa.

Autentifikacija koristi JWT, koji se čuva u skladištu sesije (engl. *session storage*) pregledača, koje traje dok je kartica otvorena. Uloge (nastavnik ili student) ugrađene su u JWT i na osnovu njih se sprovodi kontrola pristupa rutama. U razvojnom okruženju *Vite* server prosleđuje API pozive ka serverskom delu, dok se u produkciji statička verzija aplikacije poslužuje kroz *Nginx*.

3.2 Serverska aplikacija

Serverski deo je *Django* projekat podeljen na tri aplikacije, u skladu sa *Django* konvencijom razdvajanja odgovornosti:

- **Aplikacija za autentifikaciju** implementira prilagođeni korisnički model sa ulogama student i nastavnik, izdavanje i osvežavanje JWT-a, kao i klase dozvola (engl. *permission classes*) kojima se pojedinačni API pozivi ograničavaju na odgovarajuću ulogu.
- **Glavna aplikacija** definiše domenske modele, kurs, pitanje, test, studentski test i studentski odgovor, zajedno sa serijalizatorima i CRUD *endpoint*-ima, adresama REST API-ja preko kojih se izvode pojedinačne operacije nad njima. Studentski odgovor čuva i ocenu koju je dodelio nastavnik, što omogućava kasnije poređenje sa ocenama modela.
- **Aplikacija za ocenjivanje** sadrži logiku ocenjivanja i integracije sa jezičkim modelima. Za svaki ocenjeni odgovor kreira zapis sa korišćenim promptom, instrukcijama, kompletnim odgovorom modela i izdvojenom numeričkom ocenom, pa je svaka ocena naknadno proverljiva.

Sistem je zapakovan u *Docker* kontejnere: jedan kontejner izvršava *Django* aplikaciju, a drugi *Nginx* koji poslužuje klijentsku aplikaciju i prosleđuje API saobraćaj.

3.3 Integracija jezičkih modela

Jezičkim modelima pristupa se isključivo preko API-ja provajdera u oblaku; nijedan model se ne izvršava lokalno. Više provajdera pokriveno je projektnim obrascem fabrika (engl. *factory*), sa apstraktnim interfejsom *LLMIntegration* i

metodom `prompt()` čiji su parametri ulazni tekst, identifikator modela, sistemske instrukcije i temperatura. Metoda vraća par: izdvojeni tekstualni odgovor i kompletan odgovor API-ja, koji se čuva radi sledivosti.

Implementirane su tri integracije, kao tri implementacije zajedničkog interfejsa (LLMIntegration), dostupne istovremeno i birane po simboličkom imenu:

- **OpenAI**, koristi zvanični OpenAI *Python* SDK (engl. *software development kit*). Podržani modeli: `gpt-4o-mini`, `gpt-4o` i `gpt-5`.
- **DeepSeek**, koristi OpenAI-kompatibilni API *endpoint* preko istog SDK-a, sa promenjenom baznom adresom. Podržani modeli: `deepseek-chat (V3)` i `deepseek-reasoner (R1)`.
- **Google Gemini**, koristi zvanični Google GenAI *Python* SDK. Podrazumevani model: `gemini-2.5-flash`. U aplikaciji služi za generisanje simuliranih studentskih odgovora pri testiranju i nije korišćena u eksperimentima opisanim u ovom radu.

Fabrička funkcija instancira odgovarajuću klasu na osnovu simboličkog imena integracije (`openai`, `deepseek`, `gemini`). Nov provajder se tako dodaje implementacijom jedne klase, bez izmena u logici ocenjivanja; isti mehanizam koriste i veb aplikacija i sistem za automatizovanu evaluaciju. API ključevi čuvaju se kao promenljive okruženja i nikada nisu ugrađeni u izvorni kod.

Glava 4

Sistem za automatizovanu evaluaciju modela

Za potrebe evaluacije opisane u narednom poglavlju razvijen je zaseban sistem za automatizovanu evaluaciju modela, organizovan kao evaluacioni postupak (engl. *pipeline*, po čemu je nazvan i direktorijum *pipeline/*) u sistemu DVC. U DVC-u se faze obrade opisuju deklarativno, sa eksplicitnim zavisnostima i izlazima, i svaka se ponovo izvršava samo ako su se njeni ulazi ili parametri promenili. Eksperiment se tako reprodukuje bez ponovnog plaćanja API poziva čiji su ulazi ostali isti.

4.1 Format ulaznih podataka

Podaci su podeljeni u devet skupova, po jedan za svaki ispitni rok, pri čemu pojedini od pet kurseva obuhvataju više rokova. Svaki skup opisan je dvema CSV datotekama:

- `questions.csv`, po jedan red za svako pitanje, sa kolonama: identifikator pitanja, tekst pitanja, referentni odgovor nastavnika i adresa nastavnog materijala (PDF udžbenika) iz koga je pitanje postavljeno; datoteka sadrži i kolonu sa napomenom nastavnika za ocenjivanje i pomoćni identifikator, koje faze obrade ne čitaju;
- `student_answers.csv`, po jedan red za svaki studentski odgovor, sa kolonama: pseudonimni identifikator studenta, identifikator pitanja, ocena nastavnika (0–10) i tekst odgovora.

4.2 Faze obrade

Prva faza (*spajanje ulaza*) validira i spaja ove dve datoteke u jedinstvenu tabelu. Validacija obuhvata proveru obaveznih kolona u obe datoteke; za pitanja se proverava jedinstvenost identifikatora i nepraznost svih korišćenih polja, a za studentske odgovore da se pozivaju na postojeće pitanje i da svaki student ima odgovor na svako pitanje. U suprotnom se obrada prekida sa opisom greške. Provere su ovako stroge zato što su podaci pripremani ručno, a greška u njima otkrivena tek posle ocenjivanja znači uzalud utrošene resurse modela (plaćene pozive API-ja).

Druga faza (*ocenjivanje*) za svaki red spojene tabele konstruiše prompt i poziva izabrani jezički model preko integracionog sloja iz prethodnog poglavlja. Postoje tri varijante ove faze, od kojih dve odgovaraju režimima ocenjivanja PRA i GRA:

- **PRA varijanta** u prompt uključuje pitanje, referentni odgovor nastavnika i studentski odgovor.
- **GRA varijanta** umesto referentnog odgovora nastavnika koristi generisani referentni odgovor, koji model generiše iz nastavnog materijala. Ovo je dvo-stepen postupak: u prvom koraku materijal se preuzima sa zadate adrese kao PDF, tekst se izdvaja bibliotekom *PyPDF2* i kešira na disku (po heš vrednosti adrese, tako da se svaki udžbenik preuzima i obrađuje samo jednom), a model iz njega konstruiše koncizan referentni odgovor dužine dve do pet rečenica. U drugom koraku studentski odgovor ocenjuje se u odnosu na taj generisani standard, ponovnom upotrebom iste komponente kao u PRA varijanti, nad tabelom pitanja u kojoj su nastavnički referentni odgovori zamenjeni generisanim.
- **Varijanta sa nastavnim materijalom u promptu** umesto referentnog odgovora u svaki prompt za ocenjivanje umeće ceo izvučeni tekst udžbenika, istom komponentom za preuzimanje i keširanje kao GRA varijanta. Ova varijanta realizuje treći, probni režim, u kome se referentni odgovor uopšte ne koristi (odeljak 5.6).

U svim varijantama sistemska instrukcija se sastoji od prompta strogosti (blag, neutralan ili strog; dodatak A) i fiksnog uputstva o formatu izlaza: prvi red odgovora mora sadržati samo ocenu na skali 0–100, a ostatak obrazloženje. Parametarske datoteke modela zadaju temperaturu 0, ali je integracioni sloj za OpenAI

izostavlja iz zahteva kada je jednaka nuli, pa modeli porodice GPT zapravo rade na podrazumevanoj vrednosti svog provajdera, dok se modelima DeepSeek vrednost 0 uvek prosleđuje. Temperatura se ni u jednom slučaju ne varira kroz konfiguracije; njen uticaj ispitan je odvojeno, u dodatnom eksperimentu. Numerička ocena se izdvaja iz prvog reda odgovora; skala 0–100 naknadno se preslikava na skalu 0–10 radi poređenja sa ocenama nastavnika. Ako izdvajanje ne uspe, poziv se ponavlja, do tri pokušaja ukupno; ako API poziv privremeno otkáže, ponavljanju prethodi rastuće odlaganje. Redovi za koje ni posle tri pokušaja nije dobijena ocena u izlaznoj datoteci dobijaju vrednost -1 . U sprovedenim eksperimentima to je retko: u PRA i GRA režimu ocena nedostaje za 29 od 17 306 ocenjenih redova (0,17%), od čega 28 u jednom pokretanju (DeepSeek-Reasoner, neutralna konfiguracija, jedan rok kursa Programiranje 2), dakle reč je o privremenom prekidu API-ja, a ne o tome da model odbija da oceni. U režimu sa nastavnim materijalom nedostaje 237 od 3 033 ocena (7,8%), gotovo isključivo u nedovršnim pokretanjima modela GPT-4o. Pošto je skala ocenjivanja 0–100, negativna vrednost ne može biti stvarna ocena, pa je analitička skripta po njoj prepoznaje i takve redove isključuje iz izračunavanja metrika.

4.3 Automatizacija eksperimenata i analiza rezultata

Potpuni eksperiment obuhvata petnaest kombinacija (pet modela $\times$ tri konfiguracije strogosti) nad svih devet skupova podataka, dakle 135 pokretanja po režimu ocenjivanja. Pomoćna skripta za svaku kombinaciju upisuje izbor modela i strogosti u parametarsku datoteku i ponovo pokreće DVC postupak. Ime svake izlazne datoteke kodira konfiguraciju (strogost, model i režim) kojom je proizvedena. Svaka kombinacija pokrenuta je jednom, pa rezultati ne obuhvataju procenu varijabilnosti između ponovljenih pokretanja istog modela. Dva dodatna eksperimenta izvan ove mreže, sa nastavnim materijalom u promptu i sa različitim vrednostima temperature, prikazana su u odeljku 6.4.

Završna analitička skripta prolazi kroz sve ocenjene datoteke, uparuje ocene modela sa ocenama nastavnika i računa četiri metrike: PCC, odnos prosečnih ocena, koren srednje kvadratne greške (RMSE) i odnos RMSE/STDDEV, po svakoj kombinaciji skupa podataka, modela, strogosti i režima. Rezultati se objedinjuju u zbirnu tabelu iz koje su izvedeni rezultati prikazani u poglavlju 6.

4.4 Reprodukcija eksperimenata

Izvorni kod sistema, pitanja sa referentnim odgovorima nastavnika i sve agregatne metrike javno su dostupni u repozitorijumu projekta [9]. Studentski odgovori i ocenjene datoteke koje ih sadrže nisu deo javnog repozitorijuma, radi zaštite podataka studenata, već se u pseudonimizovanom obliku ustupaju na zahtev, uz saglasnost nastavnika koji su ih obezbedili; sa njima se svaki opisani eksperiment može ponoviti istim komandama. Struktura dela repozitorijuma koji se odnosi na evaluaciju prikazana je u nastavku; datoteke označene zvezdicom nisu deo javnog repozitorijuma.

Organizacija evaluacione infrastrukture

```
pipeline/
  dvc.yaml
  params.yaml
  params/models/*.yaml
  params/prompts/strictness/*.yaml
  src/pipeline/01_merge_inputs.py
  src/pipeline/02_grade_with_correct_answer.py
  src/pipeline/03_grade_with_textbook.py
  src/pipeline/04_generate_answers.py
  src/common/grading.py
  src/common/pdf_extraction.py
  run_all_models.py
  calculate_correlations.py
  calculate_temperature.py
  calculate_lengths.py
  plot_scatter.py
  data/<skup>/raw/questions.csv
  data/<skup>/raw/student_answers.csv          *
  data/<skup>/graded/graded.*.csv              *
  data/<skup>/graded/correlation_results.csv
  data/correlation_summary.csv
  data/temperature_summary.csv
  data/length_summary.csv
  data/<rok>/results/gpt-4o.default.t-*.csv    *
```

Sve datoteke navedene sa uvlačenjem nalaze se unutar direktorijuma `pipeline/`, dok se poslednja stavka odnosi na direktorijum `data/` u korenu repozitorijuma, u kome se čuvaju rezultati dodatnog eksperimenta sa temperaturom.

Datoteka `dvc.yaml` opisuje faze obrade sa njihovim zavisnostima i izlazima, a `params.yaml` sadrži spisak skupova podataka i izbor modela i konfiguracije strogosti za tekuće pokretanje. Svaki model opisan je zasebnom datotekom u `params/models`, koja navodi provajdera i tačan identifikator modela. Datoteke u `params/prompts/strictness` opisuju konfiguracije strogosti, a svaka od njih sadrži tekst sistemske instrukcije u obe varijante, sa zadatim referentnim odgovorom i sa nastavnim materijalom.

U direktorijumu `raw` nalaze se samo dve ručno pripremljene ulazne datoteke po skupu podataka, opisane na početku ovog poglavlja, dok direktorijum `graded` sadrži izlaze ocenjivanja, po jednu datoteku za svaku pokrenutu kombinaciju modela, strogosti i režima. Imena tih datoteka prate tri obrasca:

- `graded.<strogost>.<model>.correct.csv` za PRA režim;
- `graded.generated.<strogost>.<model>.csv` za GRA režim;
- `graded.<strogost>.<model>.textbook.csv` za varijantu sa nastavnim materijalom u promptu.

Oznaka `generated` stoji pre konfiguracije strogosti, a oznake `correct` i `textbook` posle imena modela. Analitička skripta iz ovih imena rekonstruiše konfiguraciju svake datoteke, pa dodatna evidencija o pokrenutim eksperimentima nije potrebna.

Za pokretanje ocenjivanja potreban je *Python 3* sa zavisnostima navedenim u datoteci `requirements.txt` i pristupni ključevi za korišćene provajdere. Ključevi se zadaju kroz dve promenljive okruženja:

- `OPENAI_API_KEY`, za modele porodice GPT;
- `DEEPSEEK_API_KEY`, za DeepSeek modele.

Sistem ih čita iz datoteke `.env.dev` u korenu repozitorijuma, koja se ne čuva u sistemu za kontrolu verzija. Same komande pokreću se iz direktorijuma `pipeline`.

Pokretanje

```
dvc repro
python run_all_models.py
python calculate_correlations.py
```

Prva komanda pokreće faze obrade za konfiguraciju zapisanu u parametarskoj datoteci, uz preskakanje faza čiji se ulazi nisu promenili. Druga komanda sprovodi ceo eksperiment tako što redom upisuje svaku od petnaest kombinacija u tu datoteku i ponovo pokreće postupak. Treća komanda prolazi kroz ocenjene datoteke svih skupova podataka, za svaki skup zapisuje metrike u njegov direktorijum `graded`, a sve rezultate objedinjuje u zbirnu tabelu `correlation_summary.csv`, čije su kolone sledeće:

Kolone zbirne tabele

```
dataset_name, model_name, strictness_level,
method, correlation, average_professor_score,
average_llm_score, average_score_ratio, rmse,
professor_stddev, rmse_stddev_ratio
```

Prve četiri kolone određuju konfiguraciju, a ostale sadrže vrednosti metrika.

Pošto su ocenjene datoteke sačuvane u celini, sve prikazane vrednosti metrika dobijaju se pokretanjem samo analitičke skripte, bez ijednog poziva jezičkom modelu i bez pristupnih ključeva. Kako su rezultati zapisani po ispitnim rokovima, vrednosti po kursevima u tabelama sa rezultatima dobijene su objedinjavanjem rokova koji pripadaju istom kursu, prema tabeli 5.1.

Jedini eksperiment koji nije pokrenut kroz opisani postupak je eksperiment sa temperaturom. On je sproveden ranije, kroz upravljačku komandu same aplikacije, sa istim promptom i istim formatom izlaza, pa se njegove ocenjene datoteke nalaze izvan direktorijuma `pipeline` i imenovane su po ispitnim rokovima iz baze aplikacije (`p2` i `p2_jul_2` odgovaraju skupovima `p2_1` i `p2_2`, a `uu` i `dzp` zovu se isto u oba slučaja). Skripta `calculate_temperature.py` uparuje te datoteke sa odgovarajućim pokretanjem iz evaluacionog postupka i računa iste metrike, pa se tabela 6.6 dobija jednom komandom, takođe bez ijednog poziva jezičkom modelu:

Reprodukcija eksperimenta sa temperaturom

```
python calculate_temperature.py
```

Pošto dva izvora koriste različite identifikatore studenata, uparivanje se sprovodi po redosledu studenata i pitanja, uz proveru poklapanja nastavničkih ocena red po red.

Na isti način, skripta `calculate_lengths.py` iz ocenjenih datoteka prebrojava prosečne dužine referentnih odgovora nastavnika, studentskih odgovora i generisanih referentnih odgovora svakog modela i zapisuje ih u `length_summary.csv`, odakle je preuzeta tabela 5.2; redovi kod kojih generisanje referentnog odgovora nije uspelo (odeljak 5.5) izostavljaju se iz proseka. Slika 6.1 dobija se skriptom `plot_scatter.py`, koja iz ocenjenih datoteka kursa Programiranje 2 čita parove ocena i crta dijagram rasejanja, sa fiksiranim semenom slučajnog pomeraja radi ponovljivosti.

Glava 5

Metodologija evaluacije

EduGrader uključuje velike jezičke modele u tok provere znanja pod nadzorom nastavnika, sa ciljem efikasnijeg i doslednijeg ocenjivanja uz očuvanje pedagoške transparentnosti i ljudske odgovornosti. Poglavlje najpre opisuje podatke i njihovu zaštitu, zatim metrike, a potom modele, konfiguracije strogosti i režime ocenjivanja, uključujući dva odstupanja od planiranog eksperimenta.

5.1 Prikupljanje podataka

Evaluacija je sprovedena na autentičnim odgovorima sa teorijskih ispita, prikupljenim tokom više ispitnih rokova na dva univerziteta u Srbiji.

Skup podataka obuhvata **pet univerzitetskih kurseva** sa osnovnih i master studija, sa mešavinom konceptualnih pitanja i zadataka vezanih za programiranje:

- Programiranje 2 (P2): kurs prve godine osnovnih studija, Matematički fakultet, Univerzitet u Beogradu;
- Tehničko i naučno pisanje (TNP): kurs prve godine osnovnih studija, Matematički fakultet, Univerzitet u Beogradu;
- Veb programiranje (VP): kurs četvrte godine osnovnih studija, Matematički fakultet, Univerzitet u Beogradu;
- Digitalni zapis podataka (DZP): kurs master studija, Univerzitet umetnosti;
- HTML5 i JavaScript za video-igre (HTML5): kurs master studija, Univerzitet umetnosti.

Skraćenice kurseva u zagradama izvedene su iz srpskih imena kurseva i poklapaju se sa imenima skupova podataka u repozitorijumu (p2, tnp, vp, dzp); skup kursa HTML5 nosi ime uu, po instituciji na kojoj se kurs drži. U objavljenom radu [11] korišćene su engleske skraćenice kurseva, pa se tabele sa rezultatima u ovom radu od tabela u članku razlikuju samo po oznakama kurseva.

Podaci su organizovani po ispitnim rokovima, a ne po kursovima: postoji po jedan skup za svaki od devet rokova, koji se grupišu u pet navedenih kurseva (tri roka za Programiranje 2, po dva za Tehničko i naučno pisanje i Veb programiranje, i po jedan za Digitalni zapis podataka i HTML5). Svi eksperimenti izvršavaju se nad pojedinačnim rokovima, dok se rezultati prikazuju objedinjeno i po kursovima.

Uz svaki kurs, skup podataka sadrži studentske odgovore, ocene koje su dodelili nastavnici, referentne odgovore nastavnika i adrese nastavnih materijala korišćenih u pripremi ispita. Za deo pitanja nastavnici su priložili i kratke napomene za ocenjivanje; one su sačuvane u ulaznim datotekama, ali ih sistem za evaluaciju ne koristi, pa u eksperimentima nisu učestvovala. Ukupno, skup sadrži **583 studentska odgovora** iz različitih disciplina (tabela 5.1). U zavisnosti od kursa, odgovori obuhvataju kratka tekstualna objašnjenja i odgovore u obliku koda, ali se tip pitanja u podacima ne beleži, pa se broj odgovora daje samo po kursu.

Tabela 5.1: Struktura skupa podataka po kursovima (broj studentskih odgovora)

Kurs	Odgovori
Programiranje 2 (P2)	477
Tehničko i naučno pisanje (TNP)	40
Veb programiranje (VP)	46
Digitalni zapis podataka (DZP)	10
HTML5 i JavaScript za video-igre (HTML5)	10
Ukupno	583

Praćena je i prosečna dužina referentnih odgovora nastavnika, generisanih referentnih odgovora svakog modela i studentskih odgovora (tabela 5.2), jer dužina referentnog odgovora utiče na to koliko studentski odgovor u poređenju sa njim deluje potpuno. Referentni odgovori nastavnika i studentski odgovori su podjednako kratki, oko 21 do 22 reči u proseku. Generisani referentni odgovori su dva do tri puta duži, od oko 47 do 67 reči u zavisnosti od modela, što odgovara traženom obimu od dve do pet rečenica (dodatak A); instrukcija o dužini je, dakle, uglavnom ispoštovana. Razlika potiče otud što na više kurseva nastavnici očekuju sažet

odgovor (npr. jedan izraz ili kratku konstataciju), dok jezički modeli i u okviru zadatog obima uz odgovor daju i kontekst i objašnjenje zašto je odgovor tačan.

Tabela 5.2: Prosečna dužina odgovora po izvoru. Generisani referentni odgovori odnose se na neutralnu konfiguraciju GRA režima. U zagradi je broj odgovora nad kojim je prosek izračunat; za modele DeepSeek izostavljeno je 477 odgovora sa kursa Programiranje 2, kod kojih generisanje referentnog odgovora nije uspeło (odeljak 5.5).

Izvor odgovora	reči	znakova
Referentni odgovor nastavnika (583)	21,3	132,5
Studentski odgovor (583)	22,2	131,6
Generisani, GPT-4o-mini (583)	67,1	431,9
Generisani, GPT-4o (583)	67,1	437,0
Generisani, GPT-5 (583)	54,4	344,1
Generisani, DeepSeek-Chat (106)	46,6	329,0
Generisani, DeepSeek-Reasoner (106)	52,3	370,9

5.2 Etička pitanja i zaštita podataka

Podaci opisani u prethodnom odeljku su autentični odgovori sa ispita, a ne odgovori prikupljeni za potrebe istraživanja. Ustupili su ih nastavnici koji su ispite sastavili i ocenili, sa Matematičkog fakulteta i Univerziteta umetnosti, isključivo za potrebe ovog rada. Studenti nisu posebno kontaktirani, pa se zaštita njihovih podataka oslanja na pseudonimizaciju sprovedenu pre bilo kakve obrade i na to da se iz samog skupa ne može utvrditi čiji je koji odgovor.

Skup podataka korišćen u radu sadrži samo četiri polja po odgovoru: pseudonimni identifikator studenta, identifikator pitanja, ocenu nastavnika i tekst odgovora. Imena, brojevi indeksa, adrese elektronske pošte ni bilo koji drugi identifikatori u skupu ne postoje. Pseudonimi (Student 1, Student 2, ...) dodeljeni su po redosledu pojavljivanja u okviru jednog ispitnog roka, i to za svaki rok posebno, pa se odgovori istog studenta iz dva roka ne mogu povezati. Tabela preslikavanja identiteta na pseudonime nije sačuvana. Jedina zadržana veza je ona između odgovora jednog studenta unutar istog roka; ona je potrebna validaciji ulaza (svaki student ima odgovor na svako pitanje) i ne nosi identitet.

Pored identifikatora, pregledan je i slobodan tekst: studentski odgovori, tekst pitanja, referentni odgovori i napomene nastavnika, kako pretragom po obrascima

(brojevi indeksa, adrese, brojevi telefona, potpisi) tako i ručnim čitanjem. Lični podaci u njima nisu nađeni, što je i očekivano s obzirom na to da su pitanja stručna (programiranje, računarske mreže, naučno pisanje) i da se odgovori ne odnose na studenta lično. Ovakav postupak je pseudonimizacija, a ne potpuna anonimizacija: nastavnik koji je ispit ocenjivao može prepoznati pojedini odgovor, ali treće lice iz samih podataka identitet studenta ne može utvrditi.

Studentski odgovori se, i pored pseudonimizacije, ne objavljuju javno. Javni repozitorijum projekta sadrži izvorni kod, pitanja sa referentnim odgovorima nastavnika i agregatne metrike, a datoteke sa studentskim odgovorima, ulazne i ocenjene, dostupne su na zahtev, uz saglasnost nastavnika koji su ih ustupili.

Isti princip važi za podatke koji napuštaju sistem. Ocenjivanje se izvodi preko API-ja provajdera u oblaku, a prompt koji se modelu šalje sadrži isključivo tekst pitanja, referentni odgovor i tekst studentskog odgovora, bez identifikatora studenta, i u evaluacionom postupku i u aplikaciji. Provajderima se tako prosleđuje samo sadržaj koji je već pseudonimizovan. U aplikaciji se identitet korisnika čuva isključivo u lokalnoj bazi institucije koja aplikaciju izvršava; ta baza, kao ni pristupni ključevi, nije deo javnog repozitorijuma.

Konačno, sistem je zamišljen kao alat za podršku ocenjivanju, a ne kao ocenjivač: konačna ocena i odluka o graničnim slučajevima ostaju kod nastavnika, a svaka ocena modela čuva se zajedno sa promptom i obrazloženjem, pa je naknadno proverljiva. Šira pitanja transparentnosti i odgovornosti u ocenjivanju podržanom veštačkom inteligencijom razmotrena su u pregledu literature.

5.3 Metrike i evaluacija

Ocene sistema EduGrader upoređene su sa ocenama koje su dodelila dva univerzitetska nastavnika. Svaki nastavnik je ocenio različit podskup od ukupno 583 odgovora, pri čemu su podskupovi određeni kursevima na kojima nastavnici predaju, pa se ne preklapaju. Nastavnik sa Matematičkog fakulteta ocenio je 563 odgovora (Programiranje 2: 477, Tehničko i naučno pisanje: 40, Veb programiranje: 46), a nastavnik sa Univerziteta umetnosti preostalih 20 odgovora (Digitalni zapis podataka: 10, HTML5 i JavaScript za video-igre: 10). Sistem se tako poredi sa dva ocenjivača, a ne sa jednim standardom, ali je podela izrazito neravnomerna, pa objedinjene vrednosti metrika pretežno odražavaju kriterijume nastavnika sa Matematičkog fakulteta. Svaki odgovor ocenjivan je zasebno, poenima na skali

0–10, prema kriterijumima koji se koriste na datom kursu, to jest prema konceptualnoj tačnosti i potpunosti odgovora. Nastavnici su uzimali u obzir da li je student ispravno identifikovao ključne koncepte, da li nedostaju važni elementi i da li je rezonovanje delimično tačno, ali nepotpuno. Odgovori koji demonstriraju ispravne osnovne ideje dobijali su delimične poene čak i kada nedostaju terminologija, struktura ili pojedini detalji, dok su suštinski netačni odgovori dobijali niske ocene ili nula poena. Za svaki odgovor, ocena koju je dodelio nastavnik zadužen za kurs tretirana je kao referentna ocena i upoređena sa ocenom sistema, preslikanom na istu skalu 0–10.

Korišćene su četiri metrike: Pirsonov koeficijent korelacije (PCC), odnos prosečnih ocena, koren srednje kvadratne greške (RMSE) i odnos korena srednje kvadratne greške i standardne devijacije (RMSE/STDDEV).

PCC [25] meri jačinu linearne povezanosti između ocena koje su dodelili ljudi i ocena koje je dodelio sistem. Koeficijent uzima vrednosti od -1 do $+1$, gde $+1$ označava savršenu pozitivnu linearnu vezu, 0 odsustvo linearne veze, a -1 savršenu negativnu linearnu vezu. PCC ne meri apsolutno slaganje: sistem koji sve ocene dosledno umanjuje za istu vrednost postigao bi savršenu korelaciju uz očigledno pogrešne ocene. PCC je zato dopunjen metrikama koje obuhvataju apsolutna odstupanja između ocena sistema i nastavnika. U radu se zbog toga dosledno razlikuju korelacija, koju meri PCC, i slaganje, koje mere RMSE i RMSE/STDDEV; visoka korelacija ne podrazumeva slaganje, što je poznat argument iz metodologije poređenja metoda merenja [19].

Odnos prosečnih ocena poredi prosečnu ocenu sistema sa prosečnom ocenom nastavnika i otkriva sistematsku blagost ili strogost, ali ne i usklađenost u variranju ocena.

Koren srednje kvadratne greške (RMSE) [8] obuhvata prosečnu veličinu razlika između ljudskih ocena i ocena sistema; manja vrednost RMSE ukazuje na jače slaganje. Odnos RMSE/STDDEV [20] meri koliko se ocene sistema slažu sa ljudskim ocenama u odnosu na prirodnu varijabilnost samih ljudskih ocena, odnosno, da li su odstupanja sistema od ljudskog suda mala u poređenju sa tim koliko ljudske ocene uopšte variraju. Vrednost odnosa RMSE/STDDEV blizu 0 ukazuje na veoma visoko slaganje, vrednost oko 1 znači da su odstupanja sistema uporediva sa tipičnom varijabilnošću ljudskog ocenjivanja, dok vrednost veća od 1 sugerise da se ocene sistema razlikuju od ljudskih više nego što se ljudi razlikuju međusobno.

5.4 Modeli i konfiguracije

EduGrader koristi višemodelski pristup ocenjivanju koji uključuje i modele opšte namene i modele za rezonovanje:

- modeli opšte namene: GPT-4o, GPT-4o-mini, DeepSeek-Chat;
- modeli za rezonovanje: GPT-5, DeepSeek-Reasoner.

Svaki model evaluiran je u tri konfiguracije strogosti ocenjivanja: blagoj, neutralnoj i strogoj. Konfiguracije se razlikuju samo u promptu, koji oponaša različite filozofije ocenjivanja iz akademske prakse. Strogost, dakle, nije numerički parametar: zadaje se isključivo formulacijom sistemske instrukcije i oslanja se na to kako model protumači pojmove blagog odnosno strogog ocenjivača. Stroga konfiguracija predstavlja model kao „strogog profesora visokoakademske institucije”, usmerava mu pažnju na stavke koje se ne poklapaju sa studentskim odgovorom i zadaje eksplicitnu meru kazne (oduzimanje petine poena po nepoklapanju), dok blaga konfiguracija usmerava pažnju na stavke koje se poklapaju. Neutralna konfiguracija ne uvodi nikakvo dodatno pravilo i predstavlja osnovnu postavku ocenjivanja. Koliko dosledno modeli razlikuju nivoe vidi se iz odnosa prosečnih ocena u tabeli 6.1: razlike među konfiguracijama su male i nisu monotone kod svih modela. Puni tekstovi svih promptova navedeni su u dodatku A. Podela na tri konfiguracije strogosti prati nalaz prethodnih studija da i ljudski ocenjivači variraju u strogosti [13].

Ocenjivanje je izvedeno u dva režima opisana u uvodu: PRA sa referentnim odgovorom nastavnika i GRA sa referentnim odgovorom koji generiše sam model. PRA režim pokriva ispite za koje postoji referentni odgovor nastavnika, a GRA one na kojima ga nastavnik ne obezbeđuje. U GRA režimu referentni odgovor se generiše iz udžbenika čija je adresa data uz pitanje, u obimu od dve do pet rečenica, a u svim pokretanjima isti model koji ga generiše zatim i ocenjuje u odnosu na njega. Modelu se pri tome prilaže ceo udžbenik, jer ispiti pokrivaju gradivo celog kursa, a pitanja nisu ručno preslikana na poglavlja, pa sistem unapred ne zna iz kog dela udžbenika pitanje potiče. Automatsko izdvajanje relevantnog odlomka (engl. *retrieval-augmented generation*, RAG) ostavljeno je za budući rad.

Mreža konfiguracija ne varira temperaturu. Temperatura je ispitana odvojeno, na podskupu odgovora i za model GPT-4o. Svi eksperimenti su na srpskom jeziku: pitanja, studentski i referentni odgovori i sistemske instrukcije.

5.5 Ograničenje GRA režima za modele DeepSeek

Pri proveru podataka posle završenih eksperimenata utvrđeno je da GRA režim za dva modela nije izveden onako kako je opisan. Nastavni materijal za kurs Programiranje 2 je ceo udžbenik, čiji izdvojeni tekst zauzima oko 255 000 tokena, jedinica teksta na koje model deli ulaz, dok kontekstni prozor (engl. *context window*) modela DeepSeek-Chat i DeepSeek-Reasoner iznosi 131 072 tokena. Integracija za OpenAI takav ulaz automatski skraćuje na veličinu kontekstnog prozora, pa su modeli porodice GPT referentne odgovore generisali iz skraćenog udžbenika. Integracija za DeepSeek tu mogućnost nema: API je svaki zahtev odbio, a faza generisanja je posle tri neuspela pokušaja u kolonu referentnog odgovora upisala tekst poruke o grešci. Faza ocenjivanja je potom ta pitanja obradila kao i sva ostala, pa je za oba modela DeepSeek, u sve tri konfiguracije strogosti, svih 477 odgovora sa kursa Programiranje 2 ocenjeno tako što je model umesto referentnog odgovora video poruku o grešci, odnosno ocenjivao je bez referentnog odgovora, iz sopstvenog znanja. Za preostala četiri kursa (106 odgovora) udžbenici su znatno manji i generisanje je uspelo.

Greška je otkrivena kada eksperimenti više nisu ponavljani, pa rezultati nisu prepravljani. Umesto toga, GRA vrednosti za modele DeepSeek prikazane su u rezultatima onako kako su dobijene, uz izričitu napomenu da na kursu Programiranje 2 ne mere ocenjivanje u odnosu na generisani referentni odgovor, već ocenjivanje bez ikakvog referentnog odgovora, a tumačenje tih vrednosti ograničeno je na ono što takva postavka dopušta. Ista okolnost važi i za tabelu 5.2, u kojoj su ti odgovori izostavljeni iz proseka dužine.

5.6 Treći režim: ocenjivanje na osnovu nastavnog materijala

Pored režima PRA i GRA, razmatran je i treći pristup, u kome se referentni odgovor uopšte ne koristi. U njemu se ceo tekst nastavnog materijala, izdvojen iz PDF udžbenika, umeće direktno u prompt zajedno sa pitanjem i studentskim odgovorom, a od modela se traži da odgovor oceni oslanjajući se isključivo na taj materijal. Ceo udžbenik se prilaže iz istih razloga kao u GRA režimu. Za

razliku od GRA režima, gde se udžbenik koristi jednom, da bi se iz njega izveo kratak referentni odgovor, ovde udžbenik ulazi u svaki pojedinačni poziv za ocenjivanje. Drugačije nije bilo moguće: API za dopunu razgovora nema stanje, pa ne postoji sesija u kojoj bi se udžbenik priložio jednom, a zatim redom ocenjivali odgovori, već svaki poziv nosi ceo kontekst. Provajderi nude jedino automatsko keširanje zajedničkog početka prompta, koje umanjuje cenu ulaznih tokena, ali je ne uklanja. U ovoj implementaciji udžbenik stoji na kraju prompta, iza pitanja i studentskog odgovora, pa se ni to nije primenjivalo. Motivacija je bila da izostavljanje posredničkog koraka spreči gubitak informacija pri generisanju referentnog odgovora, naročito kod pitanja čiji odgovor zahteva širi kontekst.

Režim je implementiran i pokrenut, ali **nije sproveden za sve modele**, pa se njegovi rezultati prikazuju odvojeno od glavnih. Razlog je cena: budući da ceo udžbenik ulazi u svaki poziv, veličina prompta raste za red veličine. Medijana broja tokena u promptu je iznosila oko 14 500 za kurs Programiranje 2, naspram oko 300 tokena u PRA režimu nad istim odgovorima, dakle približno **48 puta više**. Sprovedeni deo eksperimenta potrošio je oko 38,5 miliona tokena u promptovima i pokrio samo oko trećine planirane mreže konfiguracija. Procenjeni ukupan broj tokena za kompletnu mrežu u ovom režimu iznosi približno 119 miliona, dakle još oko 80 miliona povrh već potrošenih.

Zbog toga su evaluirani samo modeli GPT-4o-mini i GPT-4o, dok modeli za rezonovanje (GPT-5 i DeepSeek-Reasoner) i DeepSeek-Chat nisu pokretani. Ocenjivanje je u potpunosti sprovedeno jedino za GPT-4o-mini, kroz sve tri konfiguracije strogosti, dok je za GPT-4o blaga konfiguracija ostala praktično nepokrenuta, a stroga nepotpuna.

Glava 6

Rezultati

Rezultati su iskazani kroz tri od četiri metrike: PCC, odnos prosečnih ocena i odnos RMSE/STDDEV. RMSE se ne prikazuje posebno jer ulazi u odnos RMSE/STDDEV, koji istu grešku izražava u odnosu na varijabilnost ljudskih ocena i time je uporediv između kurseva sa različitim rasponom ocena. Sve vrednosti izračunate su nad objedinjenim skupom svih ocenjenih odgovora, a ne kao prosek vrednosti po pojedinačnim skupovima podataka.

Svaka konfiguracija (model $\times$ strogost $\times$ režim) pokrenuta je jednom, pa rezultati odražavaju jedno uzorkovanje iz svakog modela. Temperatura se u glavnoj mreži konfiguracija ne varira: modeli porodice GPT rade na podrazumevanoj vrednosti svog provajdera, a modelima DeepSeek prosleđuje se vrednost 0. Njen uticaj ispitan je odvojeno, na kraju ovog poglavlja.

6.1 Performanse modela

Tabela 6.1: Metrike performansi za PRA režim kroz modele i konfiguracije strogosti. Najbolji rezultati za svaku metriku i konfiguraciju strogosti označeni su podebljano. Za odnos prosečnih ocena najboljom se smatra vrednost najbliža 1,0.

Model	PCC			Odnos prosečnih ocena			RMSE/STDDEV		
	blag	neutr.	strog	blag	neutr.	strog	blag	neutr.	strog
GPT-4o-mini	0,82	0,80	0,77	0,96	0,93	0,97	0,57	0,61	0,65
GPT-4o	0,80	0,84	0,79	0,90	1,00	0,79	0,65	0,55	0,74
GPT-5	0,89	0,88	0,77	0,94	0,91	1,00	0,47	0,50	0,65
DeepSeek-Chat	0,81	0,82	0,83	0,80	0,84	0,86	0,71	0,65	0,61
DeepSeek-Reasoner	0,87	0,86	0,78	0,88	0,86	0,95	0,53	0,56	0,63

Vrednosti u tabeli 6.1 izračunate su nad svih 583 odgovora, osim u dva slučaja u kojima ocenjivanje nije uspeo za sve odgovore: neutralna konfiguracija modela GPT-4o obuhvata 399 odgovora, a neutralna konfiguracija modela DeepSeek-Reasoner 555 odgovora¹.

Modeli za rezonovanje postižu najvišu korelaciju i najmanja odstupanja od nastavnika. GPT-5 postiže najvišu korelaciju u blagoj i neutralnoj konfiguraciji ($PCC = 0,89$ i $0,88$) uz najniže odnose $RMSE/STDDEV$ ($0,47$ i $0,50$), a DeepSeek-Reasoner ga prati u stopu ($0,87$ i $0,86$). Izraženo u izvornim poenima, $RMSE$ modela GPT-5 iznosi 1,91 poen u blagoj i 2,04 u neutralnoj konfiguraciji na skali 0–10: ocena sistema tipično odstupa od nastavničke za oko dva poena, pri čemu standardna devijacija samih nastavničkih ocena iznosi 4,07 poena. GPT-4o je konkurentan u neutralnoj konfiguraciji ($PCC = 0,84$), gde postiže i odnos prosečnih ocena praktično jednak 1,0, dok su GPT-4o-mini i DeepSeek-Chat dosledno nešto slabiji.

Stroga konfiguracija menja poredak: u njoj najvišu korelaciju postiže DeepSeek-Chat ($PCC = 0,83$), a GPT-5 i GPT-4o-mini padaju na 0,77. DeepSeek-Chat je pri tom jedini model kod koga korelacija monotonno raste sa strogošću ($0,81 \rightarrow 0,82 \rightarrow 0,83$). Kod modela GPT-4o stroga konfiguracija uz to unosi odstupanje naniže (odnos prosečnih ocena 0,79), odnosno ocenjivanje znatno strože od nastavničkog.

Po odnosu prosečnih ocena modeli porodice GPT ostaju bliži nastavnicima: GPT-4o-mini se drži u rasponu 0,93–0,97, dok DeepSeek-Chat u svim konfiguracijama ocenjuje konzervativnije (0,80–0,86).

Tabela 6.2: Metrike performansi za GRA režim. Najbolji rezultati za svaku metriku i konfiguraciju strogosti označeni su podebljano. Za modele DeepSeek-Chat i DeepSeek-Reasoner generisanje referentnog odgovora nije uspeo na kursu Programiranje 2 (477 od 583 odgovora), pa njihove vrednosti pretežno odražavaju ocenjivanje bez referentnog odgovora (odjeljak 5.5).

Model	PCC			Odnos prosečnih ocena			RMSE/STDDEV		
	blag	neutr.	strog	blag	neutr.	strog	blag	neutr.	strog
GPT-4o-mini	0,63	0,65	0,61	0,79	0,77	0,82	0,85	0,84	0,84
GPT-4o	0,69	0,75	0,65	0,78	0,88	0,61	0,86	0,69	1,04
GPT-5	0,85	0,84	0,73	0,87	0,83	0,89	0,57	0,61	0,73
DeepSeek-Chat	0,43	0,50	0,47	0,37	0,52	0,42	1,42	1,22	1,34
DeepSeek-Reasoner	0,79	0,79	0,70	0,81	0,77	0,87	0,70	0,72	0,77

¹Nedostajuće ocene posledica su prekinutog pokretanja, odnosno trajno neuspešnih API poziva, koji se u izlaznim datotekama označavaju i iz analize isključuju.

U GRA režimu, gde referentni odgovor generiše model a ne nastavnik, sve metrike su lošije (tabela 6.2). Generisani referentni odgovor unosi dodatnu neizvesnost: po strukturi, formulaciji i nivou detalja može se razlikovati od odgovora koji je nastavnik imao na umu.

Među modelima porodice GPT prelazak u GRA režim najbolje podnosi GPT-5: zadržava korelaciju od 0,85 u blagoj i 0,84 u neutralnoj konfiguraciji i jedini u svim konfiguracijama drži RMSE/STDDEV ispod 0,75 (u izvornim poenima, RMSE od 2,32 u blagoj i 2,49 u neutralnoj konfiguraciji, naspram 1,91 i 2,04 u PRA režimu). Sposobniji model za rezonovanje, dakle, i bolje konstruiše referentni odgovor i bolje ocenjuje u odnosu na njega. GPT-4o u strogoj konfiguraciji prelazi granicu upotrebljivosti ($\text{RMSE/STDDEV} = 1,04$, odnos prosečnih ocena 0,61).

Vrednosti za modele DeepSeek treba čitati uz napomenu iz odeljka 5.5: na kursu Programiranje 2, koji nosi 477 od 583 odgovora, ova dva modela nisu imala generisani referentni odgovor, već su ocenjivala iz sopstvenog znanja. DeepSeek-Reasoner i u takvoj postavci zadržava korelaciju od 0,79 u blagoj i neutralnoj konfiguraciji, što govori o robusnosti modela, a ne o kvalitetu generisanog referentnog odgovora. DeepSeek-Chat bez referentnog odgovora praktično ne funkcioniše: korelacija pada na 0,43–0,50, odnos prosečnih ocena na 0,37–0,52, a RMSE/STDDEV prelazi 1,2 u svim konfiguracijama, pa njegove ocene odstupaju od nastavničkih više nego što nastavničke variraju međusobno. Koliko bi ova dva modela postigla sa stvarno generisanim referentnim odgovorom na tom kursu, ovi podaci ne pokazuju.

6.2 Poređenje PRA i GRA režima

Kroz sve modele i konfiguracije strogosti PRA režim nadmašuje GRA, po korelaciji sa ljudskim ocenama i po normalizovanim odnosima greške. Pad korelacije pri prelasku u GRA nije ujednačen: kod modela GPT-5 je najmanji (sa 0,88 na 0,84 u neutralnoj konfiguraciji), a kod modela DeepSeek-Chat najveći (sa 0,82 na 0,50), s tim što kod modela DeepSeek pad uključuje i potpuno odsustvo referentnog odgovora na najvećem kursu. Performanse u GRA režimu, dakle, presudno zavise od postojanja i kvaliteta referentnog odgovora, a ne samo od sposobnosti ocenjivanja.

Nijedna konfiguracija strogosti nije najbolja za sve modele, a optimalna se razlikuje između režima. U PRA režimu najvišu korelaciju blaga konfiguracija daje kod

modela GPT-5, GPT-4o-mini i DeepSeek-Reasoner, neutralna kod modela GPT-4o, a stroga kod modela DeepSeek-Chat. U GRA režimu neutralna je najbolja kod svih modela osim GPT-5, kod koga je to blaga, a kod modela DeepSeek-Reasoner blaga i neutralna su izjednačene (obe 0,79). Stroga konfiguracija je u oba režima najlošija za četiri od pet modela, i po korelaciji i po sistematskom odstupanju.

Sistematsko odstupanje razdvaja režime jasnije od korelacije. Dok u PRA režimu odnosi prosečnih ocena ostaju blizu 1,0, u GRA režimu svi modeli ocenjuju niže od nastavnika, u skladu sa tim što su generisani referentni odgovori po pravilu detaljniji od nastavničkih (tabela 5.2), pa studentski odgovori u odnosu na njih izgledaju nepotpuniji. Isto važi i za odnose greške: u PRA režimu svi modeli ostaju ispod 0,75, dok se u GRA režimu granica 1,0 prelazi kod modela DeepSeek-Chat u svim konfiguracijama i kod modela GPT-4o u strogoj.

Referentni odgovor nastavnika, ukupno gledano, poboljšava pouzdanost ocenjivanja kod svih modela i svih konfiguracija strogosti. GPT-5 je najjači u oba režima, a rezultat modela DeepSeek-Reasoner pokazuje da prednost modela za rezonovanje nije vezana za jednog provajdera, i da se održava čak i kada referentni odgovor izostane. Modeli opšte namene su upotrebljivi, ali osetljiviji na postavku: GPT-4o-mini nudi povoljan odnos cene i performansi u PRA režimu, GPT-4o je konkurentan u neutralnoj konfiguraciji, a DeepSeek-Chat samo u PRA režimu.

6.3 Korelacija sa ljudskim ocenama po skupovima podataka

Vrednosti PCC-a po kursevima date su u tabelama 6.3 i 6.4, u neutralnoj konfiguraciji strogosti. Budući da se broj odgovora po kursevima znatno razlikuje (tabela 5.1), vrednosti za manje kurseve treba tumačiti sa rezervom: kod kurseva DZP i HTML5 korelacija je izračunata nad po deset odgovora.

Razlika između režima najveća je na kursevima gde se očekuje kratak i precizan studentski odgovor, kao što su VP i TNP. Kod kursa TNP pad je najizraženiji, sa 0,81–0,93 na 0,46–0,72. Uzrok je to što su generisani referentni odgovori duži i opisniji od sažetih rešenja koja daje nastavnik, pa model visoko ocenjuje odgovore koji traženi postupak ispravno opisuju, ali ga ne izvršavaju. Nastavnik je takve odgovore po pravilu ocenjivao nulom. Raskorak bi se mogao ublažiti kada bi se u prompt uključila kratka napomena nastavnika o tome da li je tačno objašnjenje

dovoljno ili student mora i da izvrši traženi zadatak; sistem takav ulaz trenutno ne podržava.

Kod kurseva na kojima se očekuju duži, opisniji odgovori (DZP i HTML5) razlike između režima su znatno manje, jer generisani referentni odgovori po strukturi i nivou detalja liče na nastavničke. Na kursu HTML5 vrednosti su praktično identične u oba režima (0,93–0,97).

Najveći kurs, Programiranje 2, sa 477 od ukupno 583 odgovora, ponaša se u skladu sa objedinjenim rezultatima: GPT-5 je najbolji u oba režima (0,88 i 0,84). Na ovom kursu su upravo modeli DeepSeek ocenjivali bez referentnog odgovora, pa vrednosti 0,42 (DeepSeek-Chat) i 0,81 (DeepSeek-Reasoner) u tabeli 6.4 pokazuju koliko svaki od njih vredi kao ocenjivač bez referentnog odgovora, a ne kako bi se snašao sa generisanim referentnim odgovorom. Na četiri manja kursa, gde je generisanje uspelo, DeepSeek-Chat se drži u rasponu 0,47–0,96, uporedivo sa ostalim modelima.

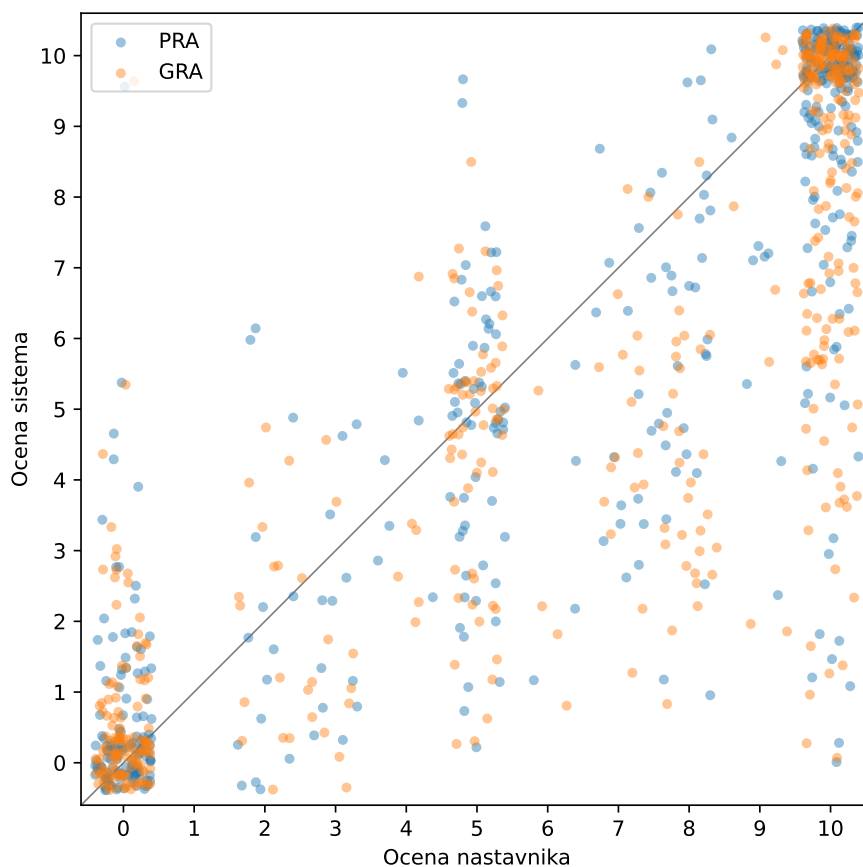
Tabela 6.3: PCC po kursevima za PRA režim (neutralna konfiguracija strogosti). U zagradi je broj odgovora nad kojim je vrednost izračunata.

Kurs	DeepSeek-Chat	DeepSeek-Reasoner	GPT-4o	GPT-4o-mini	GPT-5
P2	0,80 (477)	0,85 (449)	0,82 (293)	0,79 (477)	0,88 (477)
TNP	0,81 (40)	0,89 (40)	0,93 (40)	0,86 (40)	0,91 (40)
VP	0,78 (46)	0,74 (46)	0,81 (46)	0,78 (46)	0,75 (46)
DZP	0,80 (10)	0,82 (10)	0,61 (10)	0,86 (10)	0,92 (10)
HTML5	0,96 (10)	0,94 (10)	0,97 (10)	0,97 (10)	0,94 (10)

Tabela 6.4: PCC po kursevima za GRA režim (neutralna konfiguracija strogosti). U zagradi je broj odgovora nad kojim je vrednost izračunata. Vrednosti modela DeepSeek za kurs P2 dobijene su bez generisanog referentnog odgovora.

Kurs	DeepSeek-Chat	DeepSeek-Reasoner	GPT-4o	GPT-4o-mini	GPT-5
P2	0,42 (477)	0,81 (477)	0,74 (477)	0,64 (477)	0,84 (477)
TNP	0,47 (40)	0,46 (40)	0,72 (40)	0,62 (40)	0,65 (40)
VP	0,75 (46)	0,66 (46)	0,77 (46)	0,56 (46)	0,71 (46)
DZP	0,84 (10)	0,86 (10)	0,60 (10)	0,51 (10)	0,86 (10)
HTML5	0,96 (10)	0,94 (10)	0,96 (10)	0,97 (10)	0,93 (10)

Za najveći kurs korelacija je prikazana i vizuelno. Slika 6.1 daje dijagram rasejanja ocena na kursu Programiranje 2 za model GPT-5 u neutralnoj konfiguraciji: svaki odgovor je jedna tačka, sa ocenom nastavnika na horizontalnoj i ocenom sistema na vertikalnoj osi, a PRA i GRA režim prikazani su različitim bojama.



Slika 6.1: Dijagram rasejanja ocena nastavnika i sistema na kursu Programiranje 2 (GPT-5, neutralna konfiguracija, 477 odgovora po režimu). Svaka tačka je jedan odgovor, a siva linija označava potpuno slaganje. Tačkama je dodat slučajni pomerač do $\pm 0,4$ na obe ose, jer su ocene diskretne pa bi se inače preklapale.

Dijagram čini vrednosti PCC od 0,88 (PRA) i 0,84 (GRA) opipljivijim. Tačke se najgušće grupišu na krajevima dijagonale: nastavnik je 201 odgovor ocenio desetkom, a 125 nulom, i u tim jasnim slučajevima sistem se pretežno slaže. U PRA režimu 68% ocena sistema odstupa od nastavničke za najviše jedan poen, u GRA režimu 61%. Veća odstupanja češće idu naniže nego naviše (23% naspram 9% u PRA, odnosno 31% naspram 8% u GRA režimu), što se na dijagramu vidi kao oblak tačaka ispod dijagonale uz desnu ivicu: deo odgovora koje je nastavnik ocenio desetkom sistem ocenjuje znatno niže, izraženije u GRA režimu.

6.4 Dodatni eksperimenti

Sprovedena su i dva manja eksperimenta o odlukama koje glavna mreža konfiguracija ne pokriva: može li se posrednički referentni odgovor izostaviti i umesto njega u prompt umetnuti ceo nastavni materijal, i menja li variranje temperature slaganje sa ljudskim ocenjivanjem. Zbog ograničenog obuhvata prikazuju se odvojeno; nijedan ne menja nalaze iz prethodnih odeljaka.

Ocenjivanje iz nastavnog materijala

Režim sa nastavnim materijalom u promptu pokrenut je samo za modele GPT-4o-mini i GPT-4o, pa se nalazi odnose isključivo na ta dva modela (tabela 6.5).

Tabela 6.5: Metrike režima sa nastavnim materijalom u promptu (kolone „mat.”), uporedo sa vrednostima PRA i GRA režima za ista dva modela, prenetim iz tabela 6.1 i 6.2. Za model GPT-4o-mini sve vrednosti su nad svih 583 odgovora. Za model GPT-4o režim sa materijalom obuhvata 581 odgovor u neutralnoj i 454 u strogoj konfiguraciji, a blaga je ostala praktično nepokrenuta (12 odgovora) pa se ne prikazuje. PRA vrednosti za GPT-4o u neutralnoj konfiguraciji izračunate su nad 399 odgovora, kao u tabeli 6.1.

Model	Strogost	PCC			Odnos prosečnih ocena			RMSE/STDDEV		
		PRA	mat.	GRA	PRA	mat.	GRA	PRA	mat.	GRA
GPT-4o-mini	blaga	0,82	0,70	0,63	0,96	0,92	0,79	0,57	0,73	0,85
GPT-4o-mini	neutralna	0,80	0,70	0,65	0,93	0,92	0,77	0,61	0,73	0,84
GPT-4o-mini	stroga	0,77	0,63	0,61	0,97	0,89	0,82	0,65	0,80	0,84
GPT-4o	neutralna	0,84	0,67	0,75	1,00	0,86	0,88	0,55	0,79	0,69
GPT-4o	stroga	0,79	0,62	0,65	0,79	0,68	0,61	0,74	0,97	1,04

Kod oba modela režim sa nastavnim materijalom ostaje jasno ispod PRA režima, sa vrednostima PCC-a u rasponu 0,62–0,70 naspram 0,77–0,84, uz dosledno slabije vrednosti i po odnosu prosečnih ocena i po odnosu RMSE/STDDEV. U odnosu na GRA režim nema dosledne prednosti: kod modela GPT-4o-mini materijal u promptu daje nešto višu korelaciju u sve tri konfiguracije (0,70 naspram 0,65 u neutralnoj) i bolje slaganje po obe preostale metrike, a kod modela GPT-4o korelacija je osetno slabija (0,67 naspram 0,75), uz RMSE/STDDEV lošiji nego u GRA režimu u neutralnoj konfiguraciji.

Stroga konfiguracija se i ovde ponaša kao u GRA režimu: kod modela GPT-4o povlači odnos prosečnih ocena na 0,68, a odnos RMSE/STDDEV na granicu upotrebljivosti (0,97).

Pitanje da li izostavljanje posredničkog referentnog odgovora popravlja ocenjivanje time je delimično odgovoreno: prompt je višestruko veći nego u PRA režimu, a ni korelacija ni slaganje nisu bolji, pa dodatni kontekst ne opravdava svoju cenu. Kod modela za rezonovanje ishod bi mogao biti drugačiji, budući da oni najbolje podnose prelazak u GRA režim, ali u ovom režimu nisu pokretani.

Uticaj temperature

Temperatura upravlja slučajnošću pri uzorkovanju odgovora modela: niže vrednosti daju predvidljiviji, a više raznovrsniji izlaz. Kako se u glavnoj mreži konfiguracija ne postavlja eksplicitno, njen uticaj ispitan je posebno, za model GPT-4o u neutralnoj konfiguraciji i PRA režimu, sa temperaturama 0,5 i 1,0, nad 305 odgovora: 285 odgovora sa dva od tri roka kursa Programiranje 2 i po deset odgovora sa kurseva DZP i HTML5.

Tabela 6.6: Uticaj temperature na metrike ocenjivanja (model GPT-4o, neutralna konfiguracija strogosti, PRA režim, 305 odgovora).

Temperatura	PCC	Odnos prosečnih ocena	RMSE/STDDEV
podrazumevana	0,82	1,02	0,58
0,5	0,83	1,02	0,57
1,0	0,82	1,00	0,58

Sve tri metrike poklapaju se do druge decimale, pa variranje temperature u rasponu od 0,5 do 1,0 ne menja ni korelaciju ni slaganje sa nastavnicima na merljiv način. Red sa podrazumevanom temperaturom pri tom dolazi iz zasebnog pokretanja, a podrazumevana vrednost koju provajder primenjuje na model GPT-4o iznosi 1,0. Prvi i treći red tabele su, dakle, na istoj nominalnoj temperaturi, pa razlika između njih meri variranje između dva pokretanja istog modela, i po veličini je uporediva sa svim ostalim razlikama u tabeli.

Jedini izraženiji pomak javlja se na kursu DZP, gde PCC raste sa 0,61 na 0,89 i 0,83. I on je posledica pojedinačnih ocena među samo deset odgovora, a ne temperature. Uz ovako ograničen obuhvat, eksperiment potvrđuje samo to da izbor temperature nije prikriveni izvor razlika u glavnim rezultatima.

Glava 7

Diskusija

Rezultati iz prethodnog poglavlja ovde se stavljaju u širi kontekst. Prvo se postignute vrednosti PCC-a porede sa vrednostima objavljenim u srodnim studijama, zatim se razmatraju ograničenja evaluacije, podataka i samog sistema, a na kraju se navode pravci u kojima bi se sistem EduGrader mogao dalje razvijati.

7.1 Poređenje sa prethodnim istraživanjima

Tabela 7.1 smešta vrednosti PCC-a postignute u ovom istraživanju u kontekst prethodnih studija, čiji je pregled po dimenzijama eksperimenta dat u tabeli 2.1.

Tabela 7.1: Poređenje sa prethodno objavljenim rezultatima

Pristup	Model	PCC
Jukjevič [13]	GPT-3.5 Turbo	0,81
Kuli i Jusuf [15]	ChatGPT	0,45
Lundgren [18]	GPT-4	0,43
Naš pristup	GPT-5	0,89

Najbolja konfiguracija (GPT-5, blaga strogost) postiže $PCC = 0,89$, iznad vrednosti koju je prijavio Jukjevič (0,81, uz GPT-3.5 Turbo za ocenjivanje programerskih zadataka) [13] i daleko iznad korelacija iz studija koje su koristile ChatGPT ili GPT-4 za ocenjivanje odgovora otvorenog tipa i eseja ($PCC = 0,43$ – $0,45$) [15, 18]. Vrednost 0,43 kod Lundgrena jeste Pirsonov koeficijent korelacije, i to najviša od četiri prompt konfiguracije koje autor poredi (raspon 0,26–0,43); autor uz nju izveštava i mere slaganja (procenat slaganja i Koenova kapa), koje

su niske i nisu direktno uporedive sa PCC. Poređenja treba tumačiti sa oprezom, jer se skupovi podataka i formati zadataka razlikuju, ali dva faktora verovatno doprinose višoj korelaciji u ovoj postavci. Prvo, evaluirani su noviji modeli, pre svega modeli za rezonovanje (GPT-5, DeepSeek-Reasoner), koji stabilnije zadržavaju ponašanje kroz konfiguracije strogosti i uz visok PCC daju i niske odnose RMSE/STDDEV, blizu tipične varijabilnosti ljudskog ocenjivanja. Drugo, skup podataka sastoji se prvenstveno od kratkih teorijskih pitanja iz STEM oblasti, gde se očekuju kratka objašnjenja ili mali izrazi nalik kodu, manje dvosmisleni od dugih eseja u društvenim ili političkim naukama. Nalazi se poklapaju sa prethodnim istraživanjima i po tome što su ljudski ocenjivači u proseku nešto blaži, tolerantniji prema delimično ispravnom rezonovanju i nepotpunim odgovorima [13].

7.2 Ograničenja

Ponašanje modela može se promeniti kako se API-ji razvijaju, evaluacioni skup podataka vezan je za konkretne domene, a prestrogo kažnjavanje i pogrešno tumačenje odgovora javljaju se i u najboljim konfiguracijama.

Nekoliko ograničenja odnosi se na samu postavku evaluacije:

- **Jedno pokretanje po konfiguraciji.** Svaka kombinacija modela, strogosti i režima izvršena je jednom, pa glavni rezultati ne govore o varijabilnosti između ponovljenih pokretanja istog modela. Temperatura je zadržana na podrazumevanoj vrednosti provajdera; njen uticaj ispitan je odvojeno i nije primećen, ali samo za jedan model, jednu konfiguraciju strogosti i deo skupova podataka.
- **Konfiguracije strogosti se ne razlikuju samo po strogosti.** Tri systemske instrukcije (dodatak A) pisane su nezavisno, pa se razlikuju i mimo strogosti. Blaga i stroga instrukcija dele istu strukturu i razlikuju se u opisu uloge, u tome da li se pažnja usmerava na stavke koje se poklapaju ili na one koje se ne poklapaju sa referentnim odgovorom, i po eksplicitnom pravilu kažnjavanja (petina poena po nepoklapanju), koje postoji samo u strogoj. Neutralna instrukcija je kraća, drugačije formulisana i drugačijeg redosleda informacija, a jedina je pisana sa dijakritikom. Razlike između konfiguracija zato mere zajednički efekat strogosti, formulacije i pravopisa, a kod stroge konfiguracije i pravila o oduzimanju poena, koje je verovatno glavni uzrok

pomeranja ocena naniže. Da nivoi strogosti ne deluju kao čista skala vidi se i u tabeli 6.1: razlike u odnosu prosečnih ocena među konfiguracijama su male, a kod nekih modela ni redosled nivoa nije očekivan. Poređenje blage i stroge konfiguracije je čistije od poređenja bilo koje od njih sa neutralnom. Instrukcije nisu naknadno ujednačene, jer bi tekst u dodatku prestao da odgovara promptovima koji su dali prikazane rezultate; ponavljanje svih pokretanja sa ujednačenim instrukcijama ostavljeno je za budući rad.

- **Nepotpuni treći režim.** Ocenjivanje sa nastavnim materijalom neposredno u promptu pokrenuto je samo za dva modela, jer je veličina prompta činila potpuni eksperiment nesrazmerno skupim. Kod ta dva modela izostavljanje posredničkog referentnog odgovora ne poboljšava slaganje sa nastavnicima, a za modele za rezonovanje režim nije pokretan, pa za njih nalaza nema.
- **GRA režim bez referentnog odgovora za modele DeepSeek.** Na kursu Programiranje 2 udžbenik premašuje kontekstni prozor modela DeepSeek, pa referentni odgovor za njih nije generisan, a ocenjivanje je izvedeno bez njega (odeljak 5.5). GRA vrednosti ta dva modela zato nisu uporedive sa vrednostima modela porodice GPT, a zaključak o prednosti modela za rezonovanje u GRA režimu oslanja se na GPT-5 i na četiri manja kursa. Rešenje je izdvajanje relevantnog odlomka udžbenika pre generisanja, o čemu je reč među pravcima budućeg rada.
- **Isti model generiše i ocenjuje u GRA režimu.** Referentni odgovor i ocena potiču od istog modela, pa se kvalitet generisanog odgovora i kvalitet ocenjivanja ne mogu razdvojiti: slabiji GRA rezultat nekog modela može poticati iz bilo kog od ta dva koraka.
- **Neujednačena veličina skupova.** Kurs Programiranje 2 nosi 477 od ukupno 583 odgovora, dok kursevi DZP i HTML5 imaju po deset. Objedinjene metrike zato pretežno odražavaju jedan kurs, a vrednosti po manjim kursevima osetljive su na pojedinačne odgovore.
- **Dva ljudska ocenjivača.** Referentne ocene dodelila su dva nastavnika, svaki na svom podskupu kurseva, pa nije bilo preklapanja koje bi omogućilo procenu pouzdanosti među ocenjivačima. Slaganje sistema poredi se sa jednim nastavnikom po odgovoru, a ne sa konsenzusom. Podskupovi su uz to

izrazito neravnomerni (563 naspram 20 odgovora), pa objedinjeni rezultati gotovo u celini mere slaganje sa jednim od dvojice nastavnika.

- **Ušteda vremena nije merena.** Evaluacija pokriva slaganje sistemskih i nastavničkih ocena, a ne efekat sistema na radni proces nastavnika: koliko vremena pregled predložene ocene i obrazloženja stvarno štedi u odnosu na samostalno ocenjivanje nije ispitano.
- **Prototipska baza podataka.** Aplikacija koristi *SQLite*, dovoljan za eksperimente i demonstraciju pri lokalnom pokretanju. Višekorisnički rad sa istovremenim upisom i transakciona pouzdanost zahtevaju serverski sistem za upravljanje bazom, na koji *Django* ORM dopušta prelazak izmenom konfiguracije.

Zbog svega toga ljudski nadzor ostaje neophodan. EduGrader je zamišljen kao pomoćni sistem koji smanjuje opterećenje nastavnika, dok odluka o dvosmislenim slučajevima i konačna ocena ostaju u rukama nastavnika.

7.3 Pravci budućeg rada

Nekoliko pravaca može dodatno unaprediti sistem EduGrader:

- **Samoprovera i probni ispiti:** omogućiti studentima da iz nastavnih materijala generišu probna pitanja i tokom semestra dobijaju formativnu povratnu informaciju.
- **Usmeravanje zasnovano na pouzdanosti:** slaganje između više jezičkih modela iskoristiti kao meru pouzdanosti ocene i slučajeve niske pouzdanosti automatski označavati za proveru nastavnika.
- **Kombinovano ocenjivanje:** kombinovati modele za rezonovanje i modele opšte namene (npr. usrednjavanjem ili većinskim glasanjem) radi stabilnijih ocena od onih koje daje pojedinačni model.
- **Ocenjivanje sa primerima (engl. *few-shot*):** pored referentnog odgovora u prompt uključiti i nekoliko već ocenjenih studentskih odgovora, kako bi model kriterijum ocenjivanja video i iz primera, a ne samo iz opisa strogości. Evaluaciona infrastruktura to dopušta dopunom prompta; dodatno

obučavanje modela (engl. *fine-tuning*) nije razmatrano, jer je skup od 583 odgovora za to premali.

- **Kontrolisane varijacije prompta:** ponoviti celu mrežu konfiguracija sa tri instrukcije koje se razlikuju samo u rečenici o strogosti, uz istu strukturu, redosled informacija i pravopis, kako bi se efekat strogosti izdvojio od efekta formulacije. Strogost bi se pri tome mogla zadati i numerički, eksplicitnom kaznom po nepoklapanju ili ciljanim odnosom prosečnih ocena u svim konfiguracijama, umesto opisnim nazivima nivoa.
- **Dužina generisanog referentnog odgovora kao parametar:** prompt za generisanje traži odgovor od dve do pet rečenica, pa su generisani odgovori dva do tri puta duži od nastavničkih (tabela 5.2). Vredelo bi ispitati da li još obimniji referentni odgovor, od desetak rečenica, ostavlja više prostora za ispravno vrednovanje alternativnih formulacija studentskih odgovora, ili dodatno pojačava pomeranje ocena naniže.
- **Ukršteni GRA dizajn:** referentni odgovor generiše jedan model, a ocenjuje drugi, za sve parove modela, čime se odvojeno odgovara na pitanja koji model bolje generiše referentni odgovor, a koji bolje ocenjuje. Sistem za evaluaciju to tehnički dopušta, jer su faze generisanja i ocenjivanja odvojene: potreban je samo zaseban parametar modela za fazu generisanja.
- **Izdvajanje relevantnog odlomka nastavnog materijala:** umesto celog udžbenika, modelu priložiti samo odlomke koji se odnose na postavljeno pitanje, a koji se automatski pronalaze po sličnosti sa tekstom pitanja (pristup RAG). Prompt u režimu sa nastavnim materijalom time bi postao višestruko kraći, a generisanje referentnog odgovora u GRA režimu ne bi zavisilo od toga da li ceo udžbenik staje u kontekstni prozor modela.
- **Proširenje domena i jezika:** evaluirati EduGrader na dugim esejskim i kreativnim zadacima i u višejezičnim okruženjima, izvan STEM disciplina sa njihovim strukturiranim odgovorima.

Glava 8

Zaključak

U ovom radu predstavljen je EduGrader, višemodelska platforma otvorenog koda za ocenjivanje pitanja otvorenog tipa uz podršku veštačke inteligencije i pod nadzorom nastavnika. Sistem objedinjuje više velikih jezičkih modela (GPT-4o, GPT-4o-mini, DeepSeek-Chat i dva modela za rezonovanje, GPT-5 i DeepSeek-Reasoner) u podesiv tok ocenjivanja koji podržava evaluaciju i u PRA i u GRA režimu. EduGrader je evaluiran na 583 autentična studentska odgovora sa pet univerzitetskih kurseva na dve institucije, koji obuhvataju kratke činjeničke odgovore, teorijska objašnjenja i izraze nalik kodu. Rezultati daju odgovore na četiri pitanja postavljena u uvodu.

Usklađenost i doslednost ocena. U PRA režimu svi evaluirani modeli postižu korelaciju sa nastavničkim ocenama od najmanje 0,77, a modeli za rezonovanje najvišu: GPT-5 dostiže PCC od 0,89 uz odnos RMSE/STDDEV od 0,47, što znači da su njegova odstupanja od nastavnika manja od polovine prirodne varijabilnosti nastavničkih ocena. Korelacija, međutim, nije ujednačena po predmetima. Najviše je tamo gde se očekuje strukturiran odgovor (HTML5, TNP, P2), a najniže na kursu Veb programiranje (0,74–0,81), uz ogradu da su vrednosti za kurseve sa po deset odgovora nepouz dane. Pošto je svaka konfiguracija pokrenuta jednom, o doslednosti između ponovljenih pokretanja rad govori samo posredno, kroz eksperiment sa temperaturom, u kome dva pokretanja istog modela daju metrike jednake do druge decimale.

Kalibracija strogosti. Konfiguracija strogosti jeste upotrebljiva poluga, ali ne u oba smera podjednako. Blaga i neutralna konfiguracija daju najbolje rezultate kod četiri od pet modela, dok stroga kod većine modela snižava korelaciju i pomera ocene naniže, kod modela GPT-4o do odnosa prosečnih ocena od 0,79. Nastavnik, dakle, može promptom da pomeri prosečnu ocenu prema očekivanjima kursa, ali stroža instrukcija po pravilu plaća cenu i u korelaciji i u slaganju sa ljudskim ocenjivačem, pa je neutralna ili blaga konfiguracija sigurnija polazna tačka.

Ocenjivanje bez referentnog odgovora nastavnika. Kada referentni odgovor ne postoji, GRA režim ga generiše iz nastavnog materijala, ali po cenu kvaliteta ocenjivanja: kod svih modela i konfiguracija metrike su slabije nego u PRA režimu. Pad je najmanji kod modela GPT-5 (sa 0,88 na 0,84 u neutralnoj konfiguraciji), pa je GRA režim sa dovoljno sposobnim modelom za rezonovanje upotrebljiva zamena kada nastavnik nema referentni odgovor. Umetanje celog udžbenika u prompt umesto generisanog odgovora ne poboljšava nijednu metriku, a prompt čini oko 48 puta većim. Za modele DeepSeek ovaj deo evaluacije ostaje otvoren, jer na najvećem kursu referentni odgovor za njih nije ni generisan (odjeljak 5.5). Zaključak je da referentni odgovor nastavnika ostaje najvredniji pojedinačni ulaz u sistem.

Kontrola nastavnika nad konačnom ocenom. EduGrader je projektovan tako da automatizacija ne preuzima odluku. Ocena modela se nastavniku prikazuje uz njegovu sopstvenu, konačnu ocenu donosi nastavnik, a za svaki ocenjeni odgovor čuvaju se prompt, odgovor modela i obrazloženje, pa je svaka ocena proverljiva. Rezultati pokazuju i zašto je to neophodno: i u najboljoj konfiguraciji model greši u graničnim slučajevima, na malim kursevima vrednosti su nestabilne, a greška u GRA režimu pokazala je da se problem u ulaznim podacima može neprimećeno proširiti kroz hiljade ocena. Efikasnost automatizacije se tako koristi za rutinski deo posla, dok odgovornost za ocenu ostaje na nastavniku.

Za evaluirane tipove odgovora, a posebno za relativno strukturirana STEM pitanja, savremeni veliki jezički modeli aproksimiraju obrasce ljudskog ocenjivanja sa odstupanjima uporedivim sa prirodnom ljudskom varijabilnošću, pod uslovom da imaju referentni odgovor nastavnika i da nastavnik zadrži poslednju reč. Glavni rezultati rada predstavljeni su na konferenciji *Artificial Intelligence SANU* 2025 [10] i prihvaćeni za objavljivanje u časopisu ComSIS [11].

Dodatak A

Šabloni promptova i primeri ocenjivanja

Dodatak sadrži šablone promptova korišćene u eksperimentima i nekoliko primera ocenjivanja. Promptovi su preneti doslovno iz konfiguracionih datoteka, sa izvornim nedoslednostima u dijakritici, jer su modeli primili takav tekst.

Napomena. Model ocenjuje na skali 0–100, a ocene se u analizi i u primerima ispod preslikavaju na skalu 0–10, radi poređenja sa ocenama nastavnika.

Uputstvo o formatu izlaza

Sledeće uputstvo dodaje se na kraj sistemske instrukcije u svim konfiguracijama strogosti i u svim režimima ocenjivanja:

Tvoj odgovor treba da bude ocena studentskog odgovora u sledećem formatu:
Prvi red treba da sadrži samo broj na skali od 0 do 100, gde je 100 najbolja ocena a 0 najgora.
Ostatak odgovora treba da sadrži obrazloženje ocene.

Promptovi strogosti za PRA režim

U PRA režimu prompt uz sistemsku instrukciju sadrži pitanje, referentni odgovor nastavnika (označen kao „tačan odgovor”) i studentski odgovor, u zasebnim blokovima.

Blaga konfiguracija

Ti si blag profesor visokog obrazovanja i treba da ocenis odgovore studenata. Dobijas pitanje, tacan odgovor kao i odgovor studenta, i treba da proveris da li je odgovor studenta adekvatan. Fokusiraj se na stavke koje se poklapaju sa studentovim odgovorom.

Neutralna (podrazumevana) konfiguracija

Ti si profesor visokog obrazovanja i treba da oceniš odgovor studenta na postavljeno pitanje. Dobijaš odgovor studenta i pitanje na koje je student odgovarao, kao i tačan odgovor.

Stroga konfiguracija

Ti si strog profesor visokoakademske institucije i treba da ocenis odgovore studenata. Dobijas pitanje, tacan odgovor kao i odgovor studenta, i treba da proveris da li je odgovor studenta adekvatan. Svako nepoklapanje se kaznjava sa oduzimanjem petine ukupnih poena. Fokusiraj se na stavke koje se ne poklapaju sa studentovim odgovorom.

Neutralna konfiguracija je namerno najkraća i ne uvodi dodatna pravila; blaga i stroga usmeravaju pažnju modela na poklapanja odnosno na nepoklapanja, a stroga zadaje i meru kazne.

Prompt za generisanje referentnog odgovora (GRA režim)

U prvom koraku GRA režima model iz nastavnog materijala konstruiše referentni odgovor sledećom instrukcijom. Korak ne zavisi od strogosti; ona se primenjuje tek pri ocenjivanju, istim promptovima kao u PRA režimu.

Ti si profesor visokog obrazovanja i treba da odgovoriš na pitanje na osnovu relevantnog sadržaja iz udžbenika. Generiši koncizan odgovor koji direktno odgovara na postavljeno pitanje. Odgovor treba da bude precizan i dužine od 2 do 5 rečenica.

Promptovi za razmatrani treći režim

Instrukcije za varijantu sa nastavnim materijalom neposredno u promptu (odjeljak 5.6) razlikuju se od PRA verzija samo u opisu priloženog materijala. Kao ilustracija navedena je neutralna verzija.

Neutralna konfiguracija, varijanta sa nastavnim materijalom

Ti si profesor visokog obrazovanja i treba da oceniš odgovor studenta na postavljeno pitanje.

Dobijaš odgovor studenta i pitanje na koje je student odgovarao, kao i udzbenik iz kojeg je pitanje postavljeno.

Reprezentativni primeri ocenjivanja

Naredni primeri preuzeti su iz rezultata ocenjivanja modelom GPT-5 u neutralnoj konfiguraciji PRA režima, sa kursa Programiranje 2. Izabrani su tako da ilustruju oba smera odstupanja od nastavničke ocene.

Primer A: sistem ocenjuje strože od nastavnika

Pitanje

Koliko tačkica ispisuje poziv $f(3)$ za narednu funkciju f ?

```
void f(unsigned n) {  
    if (n > 0) {  
        f(n-1);  
        putchar('.');  
        f(n-1);  
    }  
}
```

Referentni odgovor nastavnika

Ispisuje 7 tačkica. $T(n) = 2 \cdot T(n - 1) + 1$ sa baznim slučajem $T(0) = 0$.

Studentski odgovor

Ispisuje 6 tačaka.

$f(3)$, $3 > 0$

$f(2)$, $2 > 0$

$f(1)$, $1 > 0$

$f(0)$

 .

$f(0)$

 .

$f(1)$, $1 > 0$

$f(0)$

 .

$f(0)$

 .

$f(2)$, $2 > 0$

$f(1)$, $1 > 0$

$f(0)$

 .

$f(0)$

 .

$f(1)$, $1 > 0$

$f(0)$

 .

$f(0)$

Ocene

Nastavnik: 5/10

Sistem: 3/10

Obrazloženje sistema

Netačan je krajnji rezultat: $f(3)$ ispisuje 7 tačkica, ne 6. Ispravno je $T(n) = 2 \cdot T(n - 1) + 1$ sa $T(0) = 0$, pa je $T(1) = 1$, $T(2) = 3$, $T(3) = 7$. U vašem tragu izvršavanja propušten je barem jedan ispis (npr. kod jedne od $f(1)$ grana), što je dovelo do pogrešnog prebrojavanja. Delimični poeni za pokušaj raščlanjivanja rekurzije, ali računanje i zaključak nisu tačni.

Primer ilustruje tipičan izvor neslaganja: student je postupak razložio korektno, ali je konačan broj pogrešan. Nastavnik je pokušaj nagradio polovinom poena, dok je model veću težinu dao netačnom rezultatu.

Primer B: sistem ocenjuje blaže od nastavnika

Pitanje

Da li je pristup n -tom elementu efikasniji za niz, ili za povezanu listu? Da li je pristup n -tom elementu efikasniji za niz, ili za dinamički niz? Molim Vas obrazložite odgovor.

Referentni odgovor nastavnika

Za niz. Za niz je $O(1)$, za listu $O(n)$. Isto je. U oba slučaja je $O(1)$.

Studentski odgovor

pristup n -tom elementu je lakši za niz, posto se elementi u nizu dodaju uz numeraciju, dok se kod listi pokazuje sa liste na listu bez znanja reda, što bi značilo da bi nam trebala petlja sa brojacem n dok ne dodjemo do n tog elementa liste

Ocene

Nastavnik: 5/10

Sistem: 6/10

Obrazloženje sistema

Student je tačno objasnio prvi deo: pristup n -tom elementu je efikasniji za niz nego za povezanu listu, jer se u listi mora prolaziti element po element (linearno), dok je u nizu direktan pristup. Međutim, nije odgovorio na drugi deo pitanja (niz vs. dinamički niz). Tačan odgovor za drugi deo je da je pristup n -tom elementu isti ($O(1)$) i za statički i za dinamički niz, jer oba koriste kontinuirani blok memorije i omogućavaju indeksirani pristup.

Pitanje se sastoji od dva dela, a student je odgovorio samo na prvi. Model je propust identifikovao i naveo tačan odgovor na izostavljeni deo, ali je odgovoru ipak dodelio nešto više od polovine poena, dok je nastavnik dodelio tačno polovinu. U oba primera obrazloženje sistema je sadržajno tačno i proverljivo i kada numerička ocena odstupa od nastavničke.

Dodatak B

Upotreba velikih jezičkih modela u izradi rada

Tokom izrade ovog master rada korišćeni su veliki jezički modeli kao pomoćni alat u pisanju rada i razvoju softverskog rešenja predstavljenog u radu.

Upotreba pri pisanju rada. Veliki jezički modeli korišćeni su za jezičko i stilsko unapređivanje pojedinih delova teksta, preformulisanje rečenica radi veće jasnoće i čitljivosti, kao i za pomoć pri organizaciji i strukturiranju sadržaja. Modeli nisu korišćeni kao zamena za stručnu i naučnu literaturu. Autor je proverio korišćene izvore i preuzeo odgovornost za tačnost tvrdnji, interpretacija rezultata i zaključaka iznetih u radu.

Upotreba pri razvoju projekta. Veliki jezički modeli povremeno su korišćeni kao pomoćni alat tokom razvoja softverskog sistema, prvenstveno za konsultacije u vezi sa pojedinim implementacionim pitanjima, razjašnjavanje mogućih uzroka grešaka i predloge za manje izmene ili unapređenja postojećeg koda. Svi predlozi dobijeni na ovaj način razmotreni su, prilagođeni konkretnim potrebama sistema i provereni od strane autora. Dizajn sistema, izbor tehnologija, implementacija ključnih funkcionalnosti i testiranje razvijenog rešenja bili su odgovornost autora.

Upotreba u eksperimentalnom delu. Pored upotrebe kao pomoćnog alata tokom izrade rada i projekta, veliki jezički modeli predstavljaju i predmet istraživanja ovog rada i korišćeni su u eksperimentalnoj evaluaciji sistema EduGrader. Korišćeni modeli, način njihove primene, zadavanje upita i postupak evaluacije

detaljno su opisani u odgovarajućim poglavljima rada (poglavlja 4, 5 i 6, dodatak A).

Korišćeni alati. U procesu izrade rada i razvoja projekta korišćeni su: ChatGPT (OpenAI) za konsultacije i razmatranje ideja tokom razvoja; *GitHub Copilot*, u razvojnom okruženju *VS Code*, za dopunjavanje linija koda pri kucanju; Claude (Anthropic), kroz alat *Claude Code*, za jezičko i stilsko uređivanje teksta rada i završno sređivanje korisničkog interfejsa aplikacije.

Odgovornost autora. Veliki jezički modeli korišćeni su isključivo kao pomoćni alati. Autor je samostalno donosio odluke o sadržaju rada i dizajnu i implementaciji sistema, proveravao i prilagođavao generisane predloge i izvršio konačnu proveru teksta, programskog koda, eksperimentalnih rezultata i zaključaka rada.

Literatura

- [1] Toby Baker, Laurie Smith, and Nandra Anissa. Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges. Technical report, Nesta, London, February 2019. <https://www.nesta.org.uk/report/education-rebooted/>.
- [2] Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 26–33, 2001.
- [3] Chris Brew and Claudia Leacock. Automated short answer scoring: Principles and prospects. In *Handbook of Automated Essay Evaluation*, pages 136–152. Routledge, 2013.
- [4] Susan M Brookhart, Thomas R Guskey, Alex J Bowers, James H McMillan, Jeffrey K Smith, Lisa F Smith, Michael T Stevens, and Megan E Welsh. A century of grading research: Meaning and value in the most common educational measure. *Review of educational research*, 86(4):803–848, 2016.
- [5] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 160–167, 2008.
- [6] Cristina Conati. Intelligent tutoring systems: New challenges and directions. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 2–7, Pasadena, CA, 2009.
- [7] Chaitali Diwan, Srinath Srinivasa, Gandharv Suri, Saksham Agarwal, and Prasad Ram. AI-based learning content generation and learning pathway augmentation to increase learner engagement. *Computers and Education: Artificial Intelligence*, 4:100110, 2023.

- [8] Timothy O Hodson. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14):5481–5487, 2022.
- [9] Svetozar Iković. EduGrader: izvorni kod aplikacije i evaluacione infrastrukture. GitHub repozitorijum, 2026. <https://github.com/toswe/EduGrader>.
- [10] Svetozar Iković and Jelena Graovac. EduGrader: The AI-powered grading assistant. In *Artificial Intelligence Conference: Book of Abstracts*, pages 13–14, Belgrade, October 2025. Serbian Academy of Sciences and Arts.
- [11] Svetozar Iković and Jelena Graovac. EduGrader: The AI-powered grading assistant. *Computer Science and Information Systems (ComSIS)*, 2026. Accepted for publication.
- [12] Anders Jonsson and Gunilla Svingby. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2):130–144, 2007.
- [13] Marcin Jukiewicz. The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity*, 52:101522, 2024.
- [14] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.
- [15] Chokri Kooli and Nadia Yusuf. Transforming educational assessment: Insights into the use of ChatGPT and large language models in grading. *International Journal of Human–Computer Interaction*, 41(5):3388–3399, 2025.
- [16] Rahul Kumar. Faculty members’ use of artificial intelligence to grade student papers: A case of implications. *International Journal for Educational Integrity*, 19(1):9, 2023.
- [17] Rose Luckin, Wayne Holmes, Mark Griffiths, and Laurie B Forcier. Intelligence unleashed: An argument for AI in education. Technical report, Pearson, London, 2016. <https://discovery.ucl.ac.uk/id/eprint/1475756/>.

- [18] Magnus Lundgren. Large language models in student assessment: Comparing ChatGPT and human graders. *arXiv preprint arXiv:2406.16510*, 2024.
- [19] Mohammad Ali Mansournia, Rachel Waters, Maryam Nazemipour, Martin Bland, and Douglas G Altman. Bland-Altman methods for comparing methods of measurement and response to criticisms. *Global Epidemiology*, 3:100045, 2021.
- [20] Daniel N Moriasi, Jeffrey G Arnold, Michael W Van Liew, Ronald L Bingner, R Daren Harmel, and Tamie L Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3):885–900, 2007.
- [21] Francisco Olivos, Tobias Kamelski, and Sebastián Ascuí-Gac. Assessing instructor-AI cooperation for grading essay-type questions in an introductory sociology course. *Teaching Sociology*, 54(2):132–150, 2025.
- [22] Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. Large language models for education: Grading open-ended questions using ChatGPT. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*, pages 293–302, 2023.
- [23] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527, 2022.
- [24] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.
- [25] Philip Sedgwick. Pearson’s correlation coefficient. *BMJ*, 345:e4483, 2012.
- [26] Mark D Shermis and Jill C Burstein. *Automated essay scoring: A cross-disciplinary perspective*. Routledge, 2003.
- [27] Steven E Stemler. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, 9(4), 2004.
- [28] Zachari Swiecki, Hassan Khosravi, Guanliang Chen, Roberto Martinez-Maldonado, Jason M Lodge, Sandra Milligan, Neil Selwyn, and Dragan Gaše-

- vić. Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3:100075, 2022.
- [29] Bojan B Tomić, Anisja D Kijevčanin, Zoran V Ševarac, and Jelena M Jovanović. An AI-based approach for grading students' collaboration. *IEEE Transactions on Learning Technologies*, 16(3):292–305, 2022.
- [30] Pierpaolo Vittorini, Stefano Menini, and Sara Tonelli. An AI-based system for formative and summative assessment in data science courses. *International Journal of Artificial Intelligence in Education*, 31(2):159–185, 2021.
- [31] Elizabeth L Wetzler, Kenneth S Cassidy, Margaret J Jones, Chelsea R Frazier, Nickalous A Korbut, Chelsea M Sims, Shari S Bowen, and Michael Wood. Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology*, 52(3):298–304, 2025.
- [32] Valentyna Yuskovych-Zhukovska, Tetiana Poplavska, Oksana Diachenko, Tetiana Mishenina, Yana Topolnyk, and Roman Gurevych. Application of artificial intelligence in education. problems and opportunities for sustainable development. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 13(1Sup1):339–356, 2022.
- [33] Olaf Zawacki-Richter, Victoria I Marín, Melissa Bond, and Franziska Gouverneur. Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):39, 2019.
- [34] Xiaoming Zhai. ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students*, 29(3):42–46, 2023.