

Univerzitet u Nišu
Elektronski fakultet



Tehnologije za podršku učenju

AlfL-1 Primena LLM za ocenjivanje pitanja otvorenog tipa

Mentori:

Prof. dr Ivan Milentijević
Dr Oliver M. Vojinović

Tim:

16134 Anica Jocić
18249 Milena Mijucić
17438 Vladimir Stefanović

Niš, februar 2024.

Sadržaj

Uvod	3
Background	4
Uticaj jezika na LLM	4
Uticaj semantike jezika na LLM	4
Uticaj znanja domena koji se istražuje na LLM	4
Ocenjivanje uz pomoć LLM-a	4
Primena LLM za interaktivnu nastavu	5
Upoređivanje različitih LLM-ova	5
Nedostaci	5
Poboljšanje LLM-a	5
Bezbednost koriscenja LLM-a	6
Uticaj ljudskosti na LLM	6
Postavka rešenja	7
Domen Uvod u računarstvo	7
LLM modeli i prompt inženjering	7
Proces	7
Rezultat	7
Metodologija	8
Pitanja i odgovori	8
Implementacija rešenja	11
GPT model	11
Prompt-ovi	12
Rezultati	16
Poređenje ocena	16
Zaključak	19
Literatura	20

Uvod

Primenom velikih jezičkih modela (LLM) unapređuje se proces ocenjivanja pitanja otvorenog tipa u obrazovnom kontekstu. Ovi modeli veštačke inteligencije omogućavaju efikasnu analizu odgovora, pružajući nastavnicima objektivne parametre za evaluaciju kvaliteta znanja i kreativnosti učenika. Ova tehnologija ima potencijala da obezbedi brze i pouzdane rezultate, čime olakšava pedagoški proces i omogućava novi pristup obrazovanju.

Otvorena pitanja predstavljaju formu koja ne zahteva odgovor ograničen na predefinisane opcije. Suprotno zatvorenim pitanjima, koja obično imaju jednostavne odgovore, otvorena pitanja podstiču izražavanje dubljeg razmišljanja i kreativnosti. Ona pružaju učesnicima slobodu da iskažu svoje ideje, stavove i analize, čime se podstiče dublje razumevanje i refleksija na zadatu temu, ali su samim tim i mnogo teža za ocenjivanje.

Otvorena pitanja su sastavni deo svih oblasti obrazovanja, pa ćemo u nastavku videti studije na razne teme.

Background

U akademskoj zajednici, LLM se koriste kao alat od strane istraživača i studenata, ali postoji potreba za pojašnjenjem kako o dozvoljenom obimu sadržaja generisanog od strane AI u naučnim radovima, tako i u vezi sa podobnošću LLM-a za ocenjivanje ispita.

Uticaj jezika na LLM

Pokazalo se da većina modela bolje funkcioniše na engleskom jeziku jer za druge jezike moramo više precizirati šta nam treba da bismo dobili odgovarajuće odgovore.

Uticaj semantike jezika na LLM

Istraživanje je sprovedeno procenjujući odgovore nastavnika na ispitna pitanja kao i odgovore studenata, a zatim upoređujući ta dva. Dokazano je da ChatGpt može drastično da promeni svoju procenu na osnovu manjih promena u odgovorima, a ponekad daje niže ocene zbog gramatičkih i pravopisnih problema u odgovorima koji nisu na engleskom.

Uticaj znanja domena koji se istražuje na LLM

U ovoj studiji su bile uključene dve grupe stručnjaka iz oblasti Softverskog inženjerstva i grupa učesnika studije koji su popunjavali anketu. Studija je otkrila da su učesnici koji su završili najmanje jednu tehničku obuku postigli veći rezultat u poređenju sa onima koji nisu završili nijednu obuku. Zaključili su da Chat Gpt jeste koristan za ispravljanje otvorenih pitanja i odgovora ali njegova efikasnost može zavisiti od konteksta i vrste pitanja koja se postavlja.

Ocenjivanje uz pomoć LLM-a

Odgovori 73 studenata iz kursa 'INDA at KTH Royal Institute of Technology' je ocenjivano GPT modelom. Rezultati su bili obećavajući, GPT je imao tačnost od oko 75% kada se uporedi sa ocenama asistenata koji su ocenjivali.

Ovde predstavljamo i konverzacijskog AI tutora CAIT u svrhu pomoći učenicima u rešavanju matematičkih problema u srednjoj školi. CAIT je obučen da generiše pitanja, daje savete i ispravlja greške u matematičkim problemima.

Otkriveno je da CAIT identifikuje 60% tačnih odgovora kao tačne.

Ovaj rad predstavlja CAIT kao dokaz koncepta da se CLASS okvir može primeniti za kreiranje efikasnog robota za nastavnike matematike.

Primena LLM za interaktivnu nastavu

Ispitali su potencijalnu upotrebu generativnih LLM-a za 'dialog-based' nastavu, tj. za ocenjivanje kratkih odgovora od strane učenika i generisanje interaktivnog dijaloga u nastavi.

Prednosti:

- LLM može usmeravati razgovor i njegove faze, omogućavajući profesorima da efikasno upravljaju interakcijom
- LLM često postavlja pitanja koja učenike podstiču na razmišljanje, aludirajući na dublje razumevanje gradiva

Upoređivanje različitih LLM-ova

U ovom radu se analiziraju odgovori učenika četvrtog razreda iz matematike koristeći tri LLM: GPT-3, BLOOM, i YOU.

Upoređen je njihov učinak sa rezultatima različitih klasifikatora obučeni pomoću mašinskog učenja (ML).

Otkriveno je da LLM rade lošije od ML-a u otkrivanju nekoherentnih odgovora. Nakon detaljnijeg ispitivanja, otkriveno je da se ChatGPT model suočava sa istim izazovima.

Nedostaci

- ❖ LLM nedostaje pedagoško znanje koje poseduju profesori, što rezultuje pogrešnom shvatanju gradiva kod učenika
- ❖ LLM ponekad može proizvesti netačne ili besmislene informacije i nepouzidane odgovore
- ❖ LLM nije pogodan za nastavu u hodu (on-the-fly) zbog ograničenja u reagovanju na dinamične odgovore učenika

Poboljšanje LLM-a

U ovom radu se istražuje da li su performanse LLM (Large Language Model) mreža mogu poboljšati tako što će ih učiniti dubljim i širim.

Svaki neuron u LLM mreži ima posebnu ulogu evaluacije, neuroni međusobno sarađuju i interaguju, slično ponašanju uočenom u dubokim neuronskim mrežama.

VideDeep povećava broj LLM neurona i slojeva koji sarađuju u mreži evaluacije. Proces evaluacije prati pristup unapred, pri čemu svaki sloj prima inpute iz prethodnog sloja, omogućavajući sveobuhvatnu i temeljnu procenu.

Bezbednost korišćenja LLM-a

Ovaj rad se odnosi na SuperCLUE-Safety koji je dizajniran da bi procenio bezbednost kineskih LLM-ova. Nad 13 LLM-a na kineskom jeziku, postavljali su oko 5000 otvorenih pitanja.

Istrazivanje se bavi procenom bezbednosti, kao što su: privatnost i vlasništvo, uvredljiv i štetan sadržaj, diskriminacija, etičke dileme, nasilje i agresija, politička propaganda, bezbednost dece itd.

Zaključak je da modeli sa close-source modeli nadmašuju open-source modele u pogledu bezbednosti, dakle neophodno je poboljšati bezbednost pre korišćenja od strane velikog broja ljudi.

Ovo se zasniva na modelu kineskog jezika i ne mora nužno važiti i za modele na drugim jezicima.

Uticaj ljudskosti na LLM

LLM pri donošenju "suženog izbora" koristi principe: utilitarizma (predlaže akciju koja maksimizuje dobit i korist), dužnosti (važnost ispunjavanja vlastitih moralnih obaveza), pravde (pravednost i jednakost), autonomije (individualna slobodna i sposobnost donošenja odluka), poverenja javnosti (ako se saučestvuje u društvenom kontekstu, poverenje javnosti mora biti uključeno).

Važno je da ova etička razmatranja mogu varirati u zavisnosti od kulturnih, pravnih, društvenih faktora. Iako modeli mogu biti uključeni u etičko razmatranje, oni nikako nisu zamena za ljudsku odluku, ljudska sposobnost prosuđivanja i kritičko razmišljanje su prioritet.

Postavka rešenja

Mi ćemo po uzoru na neka već rađena istraživanja¹ koristiti slične postupke i modele u našem istraživanju.

Domen Uvod u računarstvo

Ovaj domen smo izabrali zato što daje mogućnost testiranja LLM-a na raznim vrstama i složenostima pitanja, a istovremeno su ta pitanja dovoljno jasna i obrađena, tako da nama neće biti problem da ih ocenimo i uporedimo sa rezultatima LLM-a.

LLM modeli i prompt inženjering

Istraživanje će biti sprovedeno kroz nekoliko iteracija, fokusirajući se isključivo na GPT modele. verovatno eksperimentišući sa više od jednog modela. Različiti promptovi i pristupi promptovima biće korišćeni tokom ovog istraživanja. Za implementaciju samog istraživanja koristićemo OpenAI Playground, gde ćemo uneti različite varijante sistemskih promptova, modela, pitanja i odgovora.

Proces

Proces istraživanja sastoji se od rada na određenom skupu pitanja i odgovora iz relevantnog domena. Nakon toga, rezultate ćemo analizirati upoređujući ih sa našim očekivanjima putem određene metodologije i dokumentovati kao jednu iteraciju. Ovaj proces će se ponavljati više puta, s izmenama u promptovima, modelima, skupovima pitanja i odgovora, kako bismo postigli relevantne rezultate istraživanja.

Rezultat

Rezultat istraživanja će biti procena da li su GPT modeli sposobni i spremni za samostalno i dovoljno precizno ocenjivanje zadataka otvorenog tipa u oblasti Uvoda u računarstvo, pregled rezultata dobijenih podešavanjem različitih faktora koji utiču na same rezultate, naše viđenje i naš zaključak.

¹Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT (arxiv.org)

Metodologija

Naše idejno rešenje ćemo implementirati pomoću python aplikacije ili zbog brže obrade podataka i GPT modela: gpt-3.5-turbo-1106.

Sam proces testiranja će se odvijati kroz više iteracija. Svaka iteracija će koristiti ovaj model (zbog cene) i po jedan prompt, a rezultati ocenjivanja testova će biti zabeleženi zajedno sa obrazloženjem.

Pitanja i odgovori će biti na engleskom jeziku, jer kao što smo u gore navedenim studijama opisali, jezički modeli trenutno bolje funkcionišu kada radimo na engleskom.

Pitanja i odgovori

U ovom segmentu predstavimo test sastavljen od 10 pitanja koje ćemo proslediti GPT modelu na ocenjivanje. Odlučili smo se za 10 pitanja s obzirom da je to neka vrsta neformalnog standarda za broj pitanja na testovima, a takođe ne predstavlja preveliki napor za izradu, s obzirom da bi veći broj pitanja predstavljao izazov u prikupljanju odgovora. Odgovore ćemo pružiti mi, kao i kolege studenti koji će popunjavati anketu, odnosno rešavati ova pitanja.

Pitanja se razlikuju po nivou težine, s nekim veoma jednostavnim i nekim zahtevnijim, kako bismo sagledali kako će jezički model reagovati na procenu pitanja različite složenosti i formulacije. Takođe, u nastavku ćemo obezbediti tačne odgovore prikupljene od strane stručnjaka (iz literature), na osnovu kojih će se vršiti ocenjivanje pomoću jednog od sistema promptova.

Lista pitanja i odgovori stručnjaka:

- Q: Explain the difference between Moore's and Mealy's machine.
A: Moore's machine output depends only on the present state, while Mealy's machine output depends on both the present state and input.
- Q: List two reasons why representing signed integers using the method of two's complement is superior to representation using the method of one's complement.
A: Two's complement is favored over one's complement for signed integer representation due to its unique representation of zero and simplified arithmetic operations, making addition and subtraction straightforward.
- Q: Why is discretization-quantization of the value axis done when representing sound in a computer?
A: In computer sound representation, discretization involves converting continuous analog signals into discrete digital values. This process, known as quantization, allows for efficient storage and processing.
- Q: What is the purpose of ASCII codes? How many bits represent one ASCII code? A: ASCII codes serve as a standardized character encoding system. Each ASCII code is typically represented using 7 or 8 bits, accommodating 128 or 256 characters, respectively.

- Q: List the categories of processor commands and briefly explain each category. A: Processor commands can be categorized into two main types, each serving a specific function: - Arithmetic/Logic Commands: This category encompasses commands that perform mathematical and logical operations on data stored in registers. Arithmetic commands include addition, subtraction, multiplication, and division. Logic commands include operations like AND, OR, and NOT, manipulating binary data. - Control Commands: Control commands dictate the flow of execution within the processor. They include instructions for branching, jumping, and conditional execution. Control commands manage the sequence of operations, ensuring the correct execution of program instructions.
- Q: What is the purpose of the instruction register, and what is the purpose of the counter register?
A: The instruction register holds the current instruction being executed, providing guidance to the processor. The counter register keeps track of the memory address of the next instruction to be fetched.
- Q: What is software engineering? List the basic phases in software development and briefly explain each of them.
A: Software engineering is a systematic and disciplined approach to developing, designing, testing, and maintaining software. The software development consists of several key phases: - Planning: In this initial phase, project goals, requirements, and constraints are identified. Key decisions such as project scope, budget, and timeline are established. Planning sets the foundation for the entire software development process. - Analysis: The software's functionality is defined, requirements are gathered from users and the result of this phase is a requirements specification. - Design: Design involves creating a blueprint for the software based on the requirements. It includes defining the system architecture, data structures, interfaces, and algorithms. The goal is to ensure that the software will meet the specified requirements and function effectively. - Implementation (Coding): In this phase, the actual code is written based on the design specifications. Developers follow coding standards and best practices to translate the design into executable code. It's a critical step in turning the conceptual design into a functional product. - Testing: Testing is performed to identify and fix defects in the software. It includes various testing methods such as unit testing, integration testing, and system testing. The goal is to ensure that the software meets the specified requirements and works as intended. - Deployment: Deployment involves releasing the software for use. This phase includes activities such as installation, configuration, and ensuring that the software is ready for production. Deployment may also involve user training and documentation. Additional phase is Maintenance: Maintenance is an ongoing phase where the software is updated, improved, and adapted to meet changing requirements. It includes fixing bugs, adding new features, and optimizing performance. Maintenance ensures the long-term viability and usefulness of the software.

- Q: Provide one weighted BCD code, and in this code, represent the number 873.
A: The Binary-Coded Decimal (BCD) is a positional numeral system with a base of 10 in which each decimal digit is encoded using a binary alphabet (or replaced with binary words). Representing the number 873 in BCD-8421 code (the most common code) involves the binary representation 1000 0011 0111.
- 9. Q: What is a bus, and what is its purpose? Which signals are transmitted through the bus?
A: A bus is a shared communication path to which a larger number of components are connected, and the data transmitted by one component is available for reception by all other components connected to the bus. Through the bus, three signal groups can be transmitted: data, addresses, and control signals.
- Q: Write the URL address of the page "spisak_studenata.html" located in the directory "II_godina", belonging to the domain "elfak.ni.ac.rs", and hosted on "cs".
A: A URL for the page "spisak_studenata.html" in the directory "II_godina" on the host "cs" of the domain "elfak.ni.ac.rs" could look like this: `http://cs.elfak.ni.ac.rs/II_godina/spisak_studenata.html`.

Implementacija rešenja

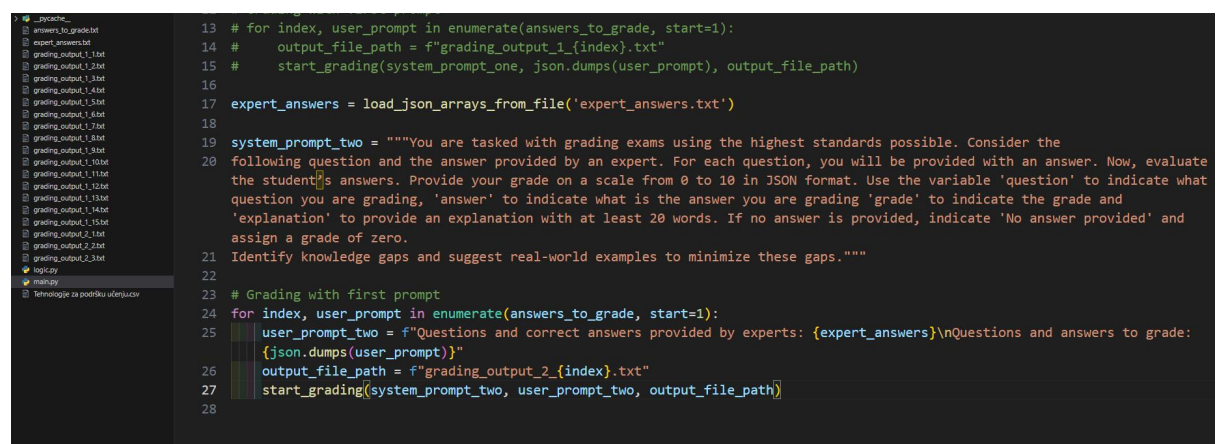
Prvo smo napravili anketu (google forms) koju je validno popunilo 15 studenata sa našeg fakulteta, na čijim odgovorima smo radili naše istraživanje. Zatim smo te odgovore mi ocenjivali, s obzirom da su pitanja bila direktna, nedvosmislena i uglavnom osnovna, to smo uradili sami. Nakon toga smo sa "google form" platforme skinuli podatke u csv formi pod nazivom 'Tehnologije za podršku učenju.csv'.

Zatim smo napravili просту konzolnu aplikaciju, koja je bila zadužena prvo za konvertovanje ovih podataka u niz validnih JSON objekata (JavaScript Object Notation) korišćenjem naših funkcija, skladištili u lokalni .txt fajl pod nazivom 'answers_to_grade.txt', a zatim ih kao takve koristili u našem user prompt-u. Onda smo te fajlove iskoristili i pokrenuli rad LLM-a i očekivane odgovore u JSON formatu pakovali u lokalne fajlove pod nazivom 'grading_output_i_j' (i predstavlja broj prompt-a iskorišćenog za generisanje, a j numerički broj odgovora (1-15)).

Nakon toga smo ove podatke obradili mi (u prvoj iteraciji samo procitali i dobili osećaj o tome koliko je validno ocenjivano, a u drugoj upoređivali sa ocenama koje smo mi dali).

GPT model

Imamo dva systemska prompt-a GPT modela koje smo koristili za ocenjivanje odgovora, koja se nalaze u nastavku teksta. Jedan je zasnovan na ocenjivanju u odnosu na znanje samog GPT modela, dok se drugi zasniva na ocenjivanju u odnosu na odgovor stručnjaka.



```
13 # for index, user_prompt in enumerate(answers_to_grade, start=1):
14 #     output_file_path = f"grading_output_1_{index}.txt"
15 #     start_grading(system_prompt_one, json.dumps(user_prompt), output_file_path)
16
17 expert_answers = load_json_arrays_from_file('expert_answers.txt')
18
19 system_prompt_two = """You are tasked with grading exams using the highest standards possible. Consider the
20 following question and the answer provided by an expert. For each question, you will be provided with an answer. Now, evaluate
21 the student's answers. Provide your grade on a scale from 0 to 10 in JSON format. Use the variable 'question' to indicate what
22 question you are grading, 'answer' to indicate what is the answer you are grading 'grade' to indicate the grade and
23 'explanation' to provide an explanation with at least 20 words. If no answer is provided, indicate 'No answer provided' and
24 assign a grade of zero.
25 Identify knowledge gaps and suggest real-world examples to minimize these gaps."""
26
27 # Grading with first prompt
28 for index, user_prompt in enumerate(answers_to_grade, start=1):
29     user_prompt_two = f"Questions and correct answers provided by experts: {expert_answers}\nQuestions and answers to grade:
30     {json.dumps(user_prompt)}"
31     output_file_path = f"grading_output_2_{index}.txt"
32     start_grading(system_prompt_two, user_prompt_two, output_file_path)
```

Koristili smo konzolni python program za parsiranje podataka iz csv-a u niz JSON objekata strukture kao na slici:

```
{
  "results": [
    {
      "question": "1. Explain the difference between Moore's and Mealy's machine.",
      "grade": 8,
      "explanation": "The answer provides a clear and accurate differentiation between Moore's and Mealy's machines. However, it could be improved by providing examples or real-world scenarios to illustrate the differences."
    },
    {
      "question": "2. List two reasons why representing signed integers using the method of two's complement is superior to representation using the method of one's complement.",
      "grade": 5,
      "explanation": "The answer mentions wider range and unambiguous representation of zero, but it lacks clarity and detail. It would benefit from explaining the overflow handling and simplification of arithmetic operations using two's complement."
    },
    {
      "question": "3. Why is discretization-quantization of the value axis done when representing sound in a computer?",
      "grade": 2,
      "explanation": "The answer is incomplete and lacks understanding of sound representation in computers. Discretization-quantization is essential to convert continuous analog sound waves into discrete digital signals for storage and processing."
    },
    {
      "question": "4. What is the purpose of ASCII codes? How many bits represent one ASCII code?",

```

Zatim smo tako parsirane podatke slali na upit gpt-u koji je koristio ovu konfiguraciju:

```
model="gpt-3.5-turbo-1106"
response_format={ "type": "json_object" }
messages=[
  {"role": "system", "content": system_prompt},
  {"role": "user", "content": user_prompt},
]
temperature=0
```

Prompt-ovi

❖ *U prvom smo koristili:*

```
user_prompt = f"Questions and answers to grade: {answers_to_grade}"
system_prompt = """You are participating in a comprehensive test covering various computer science fundamentals. Your task is to grade answers to series of questions . The questions may cover topics such as programming, algorithms, databases, and networking. For each question, you will be provided with an answer. Your response should be in JSON format, including a grade (0 to 10) and an explanation with at least 20 words. Identify any knowledge gaps and propose real-world scenarios for clarity. If no answer is provided, inform 'No answer provided', and give a grade of zero."""
```

❖ *A u drugom:*

```
user_prompt = f"Questions and correct answers provided by experts: {questions_answers}\nQuestions and answers to grade: {answers_to_grade}"
system_prompt = system_prompt_two = """You are tasked with grading exams using the highest standards possible. Consider the following question and the answer provided by an expert. For each question, you will be provided with an answer. Now, evaluate the student's answers.
```

Provide your grade on a scale from 0 to 10 in JSON format for each pair of question and answer. Use the variable 'answer' to indicate what is the answer you are grading 'grade' to indicate the grade and 'explanation' to provide an explanation with at least 20 words. If no answer is provided, indicate 'No answer provided' and assign a grade of zero. Identify knowledge gaps and suggest real-world examples to minimize these gaps.'''

Brzo smo uvideli da se javljaju isti problemi. Prvi je bio neujednačeno predstavljanje JSON izlaza pa smo zato dodali primer kako on zapravo treba da izgleda u promptu, i takođe problem gde je u nekim slučajevima sintaksne greške kažnjavao previše, pa smo unapredili prompt da vodi računa i o tome:

❖ *Prvi nakon izmene:*

```
user_prompt = f"Questions and answers to grade: {answers_to_grade}"
system_prompt_one = f'''You are participating in a comprehensive test
covering various computer science fundamentals. Your task is to grade
answers to series of questions. The questions may cover topics such as
programming, algorithms, databases, and networking. For each question, you
will be provided with an answer. Provide your grade on a scale from 0 to 10 in
JSON format for each pair of question and answer. Use the variable 'answer'
to indicate what is the answer you are grading 'grade' to indicate the grade
and 'explanation' to provide an explanation with at least 20 words like in the
example. Identify any knowledge gaps and propose real-world scenarios for
clarity. Do not punish syntax errors if the answer can be deducted. If no
answer is provided, inform 'No answer provided', and give a grade of zero.
Example of the expected output:
{json.dumps(json_output_example)}
'''
```

❖ *Drugi nakon izmene:*

```
user_prompt = f"Questions and correct answers provided by experts:
{questions_answers}\nQuestions and answers to grade: {answers_to_grade}"
system_prompt_two = f'''You are tasked with grading exams using the
highest standards possible. Consider the
following question and the answer provided by an expert. For each question,
you will be provided with an answer. Now, evaluate the student's answers.
Provide your grade on a scale from 0 to 10 in JSON format for each pair of
question and answer. Use the variable 'answer' to indicate what is the answer
you are grading 'grade' to indicate the grade and 'explanation' to provide an
explanation with at least 20 words like in the example. Do not punish syntax
errors if the answer can be deducted. If no answer is provided, indicate 'No
answer provided' and assign a grade of zero.
Identify knowledge gaps and suggest real-world examples to minimize these
gaps.
Example of the expected output:
{json.dumps(json_output_example)}
'''
```

""

```
json_output_example={
  "answers": [
    {
      "answer": "With Moore's machine the output is determined only after the
state is changed while with Mealy's output changes with input changes",
      "grade": 8,
      "explanation": "The answer provides a clear and accurate differentiation
between Moore's and Mealy's machines. However, it could be improved by
providing a more detailed explanation and real-world examples of their
applications."
    },
    {
      "answer": "Wider range of number that can be represented and zero is
represented unambiguously",
      "grade": 4,
      "explanation": "The answer correctly mentions the wider range of numbers
that can be represented using two's complement. However, it lacks clarity and
explanation of the second reason. Real-world scenarios could be provided to
illustrate the superiority of two's complement over one's complement."
    },
    {
      "answer": "I don't know",
      "grade": 0,
      "explanation": "No answer provided. The grade is zero."
    },
    {
      "answer": "ASCII is used to represent characters, it can have 7 or
extended ASCII has 8 bits",
      "grade": 7,
      "explanation": "The answer correctly identifies the purpose of ASCII codes
and the number of bits used to represent them. However, it could be improved
by providing more details about the extended ASCII and its significance in
modern computing."
    },
    {
      "answer": "Data transfer - move data Arithmetic - arithmetic operations
Logic instructions - logic operations Transfer control - manage program flow
Comparison inst - compare values Bit manipulation - modify bits Load and
store - read and write memory",
      "grade": 6,
      "explanation": "The answer provides a basic understanding of the
categories of processor commands but lacks clarity and detail in explaining
each category. Real-world examples of instructions in each category could
enhance the explanation."
    },
    {
      "answer": "Instruction register has the address of the current instruction
and counter reg has the address of the next instruction",

```

```

    "grade": 9,
    "explanation": "The answer clearly explains the purpose of the instruction
register and the counter register. It could be further improved by providing
real-world examples of how these registers are used in processor operations."
  },
  {
    "answer": "Can't remember",
    "grade": 0,
    "explanation": "No answer provided. The grade is zero."
  },
  {
    "answer": "1000 0111 0011",
    "grade": 2,
    "explanation": "The answer attempts to represent the number 873 in BCD,
but it is incorrect. The explanation lacks clarity and understanding of the BCD
system. Real-world examples of BCD representation could help in
understanding its significance."
  },
  {
    "answer": "Bus connects the components of processor. Memory bus, data
bus an control bus.",
    "grade": 7,
    "explanation": "The answer provides a basic understanding of the purpose
of a bus and mentions some of the signals transmitted through the bus.
However, it could be improved by explaining the specific role of each type of
bus and providing real-world examples of bus operations."
  },
  {
    "answer": "www.cs.elfak.ni.ac.rs/II_godina/spisak_studenata.html",
    "grade": 10,
    "explanation": "The answer correctly provides the URL address for the
given page, including the domain, directory, and page name. It is accurate
and complete."
  }
]
}

```

Rezultati

U programu postoje funkcije koje su nam pomogle u izvlačenju podataka o ocenama po pitanjima i po samim studentima u nizove, za koje smo zatim tražili srednje vrednosti, a nakon toga računali i standardnu devijaciju u odnosu na različite promptove i u odnosu na naše ocenjivanje, kao što ćemo videti kasnije.

Formula za računanje standardne devijacije:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Ona nam pokazuje zapravo odstupanja u ocenama datim određenim studentima i pitanjima. Ovo je vrlo prost metod, koji uzima mali broj faktora u obzir, pa ove vrednosti treba uzeti sa rezervom, jer veliko odstupanje (velika vrednost devijacije) ne mora značiti loše ocenjivanje, jer na samo ocenjivanje utiče ogroman broj faktora.

Poređenje ocena

Razlika između ocena promptova i naših ocena se može videti u narednim tabelama:

Tabela srednjih vrednosti po pitanjima						
Pitanje	Srednja vrednost (1)	Srednja vrednost (2)	Srednja vrednost (mi)	Std. Dev. (1) (2)	Std. Dev. (1) (mi)	Std. Dev. (2) (mi)
1	5.8	4.6	4.8	0.848528137	0.707106781	0.141421356
2	5.07	4.73	7.6	0.240416306	1.788980156	2.029396462
3	5	4.07	6.8	0.657609307	1.272792206	1.930401513
4	6.2	5.2	8.9	0.707106781	1.909188309	2.61629509
5	4.13	3.73	3.5	0.282842712	0.445477272	0.16263456
6	7.53	6.73	7.06	0.565685425	0.332340187	0.233345238
7	0.87	3.07	3.9	1.555634919	2.142533547	0.586898628
8	2	2.4	6.13	0.282842712	2.920351006	2.637508294
9	6.47	6.27	7.2	0.141421356	0.51618795	0.657609307
10	8.8	8.2	6.3	0.424264069	1.767766953	1.343502884

Tabela srednjih vrednosti po odgovorima studenata						
Pitanje	Srednja vrednost (1)	Srednja vrednost (2)	Srednja vrednost (mi)	Std. Dev. (1) (2)	Std. Dev. (1) (mi)	Std. Dev. (2) (mi)
1	5.3	5	5.1	0.212132034	0.141421356	0.070710678
2	7.2	7.4	9.2	0.141421356	1.414213562	1.272792206
3	4.9	3.8	2.7	0.777817459	1.555634919	0.777817459
4	6.7	8.5	9.1	1.272792206	1.697056275	0.424264069
5	5.6	4.9	5	0.494974747	0.424264069	0.070710678
6	6.4	6.1	9.2	0.212132034	1.979898987	2.192031022
7	5.3	5.1	7.8	0.141421356	1.767766953	1.909188309
8	6.3	6	8.5	0.212132034	1.555634919	1.767766953
9	4.9	4.2	6.2	0.494974747	0.919238816	1.414213562
10	5.2	4.8	6	0.282842712	0.565685425	0.848528137
11	4.3	4.2	5.7	0.070710678	0.989949494	1.060660172
12	4.8	4.3	6.3	0.353553391	1.060660172	1.414213562
13	5.5	4.3	7	0.848528137	1.060660172	1.909188309
14	1.9	1.7	0.9	0.141421356	0.707106781	0.565685425
15	3.5	3.2	4.6	0.212132034	0.777817459	0.989949494

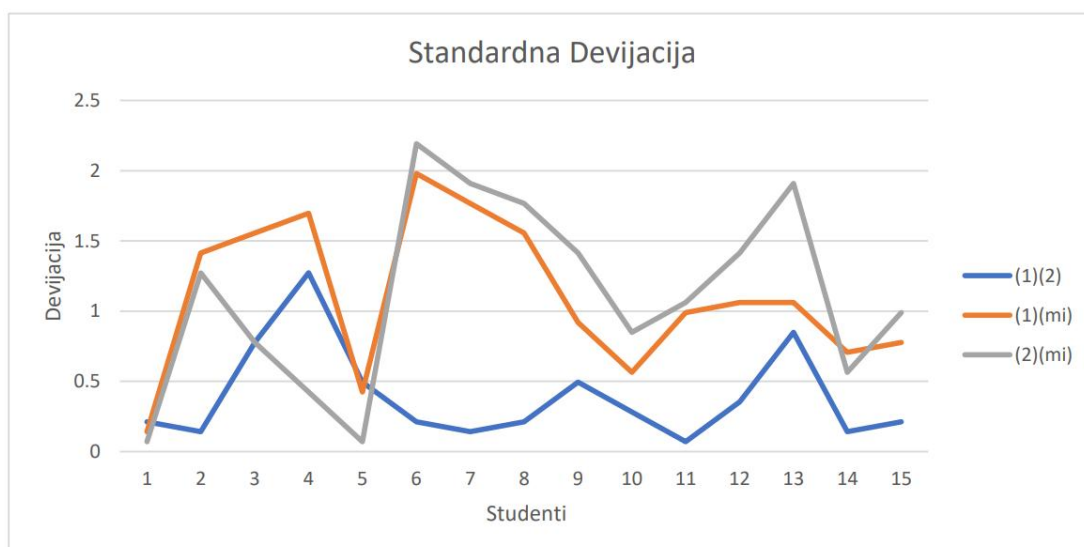
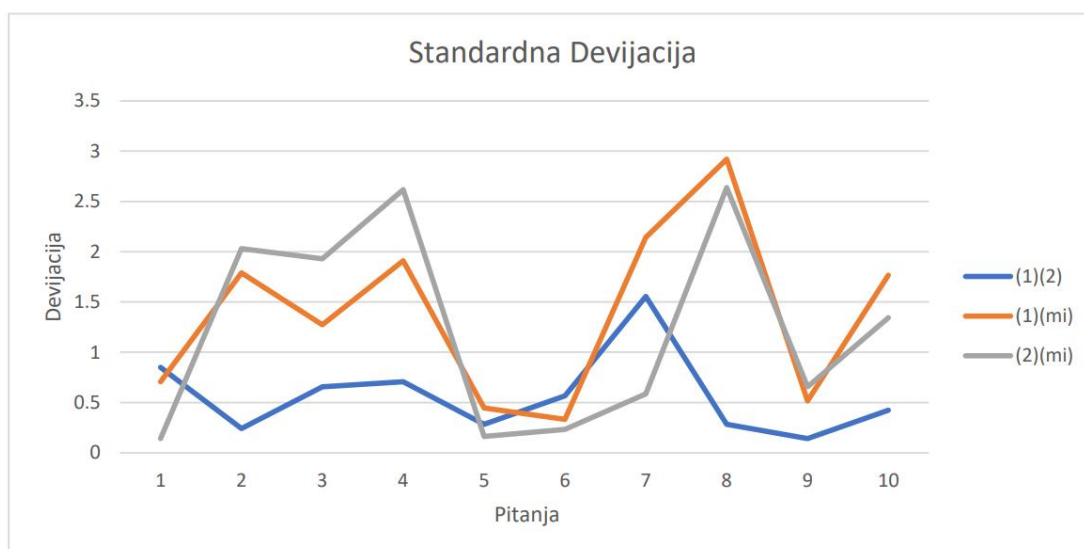
Ono što možemo uočiti na ovim slikama jeste da postoji prvenstveno razlika u ocenjivanju između samih promptova, ali i nas kao ocenjivača. Može se uočiti da kada gledamo po pitanjima, prvi LLM je davao nešto više ocene za nekih 5.37%. Takođe se može uočiti da je davao veće ocene u 80% slučajeva u odnosu na drugi. Najveće odstupanje je vezano za 7. pitanje,

“What is software engineering? List the basic phases in software development and briefly explain each of them.” gde je bio nešto stroži prvi, pa zato se tu i vidi skok u razlici.

Dalje možemo uočiti da se naše ocenjivanje u nekim pitanjima jako poklapa sa jednim ili drugim modelom, a u nekim dosta razlikuje.

U slučaju kada ga upoređujemo sa drugim promptom, možemo videti da se u 5 od 10 pitanja slažu gotovo u potpunosti, a u drugih 5 postoji dosta veća razlika.

U slučaju poređenja s prvim, slaže se sa 4 od 10, a sa 6 od 10 ne, ali ne u tako strogim granicama kao sa drugim.



Takođe, moguće je zaključiti da je GPT davao precizna objašnjenja, u malo reči kako je i navedeno u prompt-u, koja opisuju njegovo viđenje odgovora koje je dao neki student, i koja ukazuju na to zašto je ocena takva kakva jeste. Neki od primera:

```

{
    "answer": "Bus is digital circuit tasked with transmitting signals through computer. There are 3 types of buses which transmit corresponding signals: control buses, data buses, address buses.",
    "grade": 7,
    "explanation": "The answer provides a basic understanding of the purpose of a bus and mentions some of the signals transmitted through the bus. However, it could be improved by explaining the specific role of each type of bus and providing real-world examples of bus operations."
}

{
    "answer": "https://cs.elfak.ni.ac.rs/II_godina/spisak_studenata.html",
    "grade": 10,
    "explanation": "The answer correctly provides the URL address for the given page, including the domain, directory, and page name. It is accurate and complete."
}

{
    "answer": "It converts characters into computer readable code",
    "grade": 3,
    "explanation": "The answer provides a very basic understanding of the purpose of ASCII codes but lacks detail and accuracy. It fails to mention the number of bits representing one ASCII code. Real-world examples could enhance the explanation."
},

```

Zaključak

Na osnovu poređenja ocena, primetili smo nekoliko stvari:

- ❖ ako je odgovor jezički nejasan (koristi sinonime koje model ne očekuje), model ga ne razume lepo i zato mu daje manji broj poena
- ❖ promptovi (pogotovo prompt sa odgovorima stručnjaka) se drže nekih čvrstih uverenja o tome šta je tačan odgovor pa ako student da drugačiji ali ipak tačan odgovor (koji nismo očekivali i predvideli, samim tim ni naglasili promptu) oceniće odgovor sa manjim brojem poena
- ❖ promptovi daju manji broj poena jer očekuju i dodatno objašnjenje (npr. primer iz realnog sveta), dok mi najčešće samo očekujemo odgovor/definiciju (bez primera)
- ❖ npr. u poslednjem pitanju, mi kao ocenjivači gledamo da link bude jasan i precizan (cilj nam je da vidimo da li studenti znaju kako tačno treba da glasi url), dok prompt ne uzima sve to u obzir i daje maksimum iako odgovor nije baš ono što smo mi očekivali od studenata
- ❖ prompt za polovične odgovore ne daje isti broj poena kao mi (nekad više, nekad manje) iz razloga što ne ume da proceni šta je baš bitno za nas (odnosno šta tačno tražimo od studenata) pa tako za deo odgovora koji mi cenimo kao 2/10 on da 6/10 (iako fali onaj nama bitniji deo odgovora), tako da bismo morali mnogo preciznije da mu zadajemo kako da ocenjuje, morali bismo dosta dodatnih informacija da mu obezbedimo
- ❖ promptovi su među sobom takvi da prompt koji ima odgovore stručnjaka strože ocenjuje jer se više drži odgovora koje smo mu zadali kao tačne, ali nije prevelika razlika u ocenjivanju u odnosu na prompt bez unapred zadatih odgovora

Na osnovu nabrojanog, zaključujemo da se za ocenjivanje pitanja otvorenog tipa dosta uspešno može koristiti veštačka inteligencija, uz određenu dozu opreza jer je i dalje ocenjivanje na samom profesoru, odnosno on najbolje zna šta mu je prioritet pri ocenjivanju odgovora studenata. Naravno, mi možemo promptu dosta detaljno opisati na šta treba da se fokusira pri ocenjivanju i time ćemo poboljšati njegov rad i prilagoditi ga svojim kriterijumima što je više moguće.

Literatura

- <https://arxiv.org/ftp/arxiv/papers/2309/2309.11508.pdf>
- [Generative Large Language Models for Dialog-Based Tutoring: An Early Consideration of Opportunities and Concerns](#)
- <https://ceur-ws.org/Vol-3487/paper4.pdf>
- [Analyzing Multilingual Competency of LLMs in Multi-Turn Instruction Following: A Case Study of Arabic](#)
- <https://arxiv.org/pdf/2310.14819.pdf>
- [Towards LLM-based Autograding for Short Textual Answers](#)
- <https://arxiv.org/ftp/arxiv/papers/2309/2309.11508.pdf>
- [Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT](#)
- <https://arxiv.org/pdf/2307.16696.pdf>
- [GPT-4 as an Automatic Grader: The accuracy of grades set by GPT-4 on introductory programming assignments](#)
- <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1779778&dswid=4448>
- [Utilizing the CLASS Framework to Develop a Conversational AI Tutor for open-ended problems in ASSISTments](#)
- <https://par.nsf.gov/biblio/10470443>
- [A large language model-assisted education tool to provide feedback on open-ended responses](#)
- <https://arxiv.org/pdf/2308.02439>
- [Who's the Best Detective? Large Language Models vs. Traditional Machine Learning in Detecting Incoherent Fourth Grade Math Answers](#)
- <https://journals.sagepub.com/doi/abs/10.1177/07356331231191174>
- <https://ceur-ws.org/Vol-3487/paper4.pdf>
- <https://www.medrxiv.org/content/10.1101/2023.03.25.23287743v1.full.pdf>
- [Wider and Deeper LLM Networks are Fairer LLM Evaluators](#)
- <https://arxiv.org/pdf/2308.01862.pdf>

- [SC-Safety: A Multi-round Open-ended Question Adversarial Safety Benchmark for Large Language Models in Chinese](#)
- <https://arxiv.org/pdf/2310.05818.pdf>
- <https://arxiv.org/pdf/2305.11175.pdf>
- [Towards Answering Open-ended Ethical Quandary Questions](#)
- <https://amulyayadav.github.io/AI4SG2023/images/22.pdf>

Samoevaluacija studenata:

Anica Jocić 16134- prezentacija i poređenje ocena

Vladimir Stefanović 17438- aplikacija, test primeri, grafički podaci

Milena Mijucić 18249- dokumentacija i ocenjivanje s naše strane