SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

**Ivan Vlašić**

**Marko Kir**

Usporedba performansi različitih velikih jezičnih modela u alatu Promptly

Projekt

Varaždin, 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Ivan Vlašić

Marko Kir

Studij: Informacijski i poslovni sustavi

**Usporedba performansi različitih velikih jezičnih modela u alatu Promptly**

Projekt

Mentor/Mentorica:

Izv. Doc. dr. sc. Dijana Oreški

Varaždin, travanj 2024.

*Ivan Vlašić i Marko Kir*

Izjava o izvornosti

Izjavljujemo da je naš projekt izvorni rezultat našeg rada te da se u izradi istoga nismo koristili drugim izvorima osim onima koji su u njemu navedeni. Za izradu projekta su korištene etički prikladne i prihvatljive metode i tehnike rada.

*Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Sadržaj

[5. Sadržaj iv](#_Toc168131102)

[1. Uvod 1](#_Toc168131103)

[2. Razumijevanje domene 2](#_Toc168131104)

[3. Razumijevanje podataka 8](#_Toc168131106)

[4. Pregled Alata 19](#_Toc168131107)

[5. Prikaz strukture chatbot modela 20](#_Toc168131108)

[5.1. Google Gemini 20](#_Toc168131109)

[5.2. OpenAI chatbot 23](#_Toc168131110)

[6. Prikaz rada chatbota 23](#_Toc168131111)

[6.1. Razgovor s Google GEMINI 24](#_Toc168131112)

[6.2. Razgovor s OpenAI 27](#_Toc168131113)

[7. Usporedba Performansi 30](#_Toc168131114)

[8. Zaključak 33](#_Toc168131115)

[9. Popis literature 34](#_Toc168131116)

[10. Popis slika 35](#_Toc168131117)

1. Uvod

Brza i precizna obrada prirodnog jezika postaje sve važnija u različitim područjima, uključujući ali ne ograničavajući se na strojno prevođenje, generiranje teksta, pretraživanje informacija i interakciju s korisnicima putem raznih platformi. U svijetu umjetne inteligencije, veliki jezični modeli postaju ključni alati za ostvarivanje napretka u obradi prirodnog jezika. U tom kontekstu, alat Promptly predstavlja jedan od relevantnih alata koji omogućuje istraživačima i razvojnim timovima brzu i jednostavno korištenje različitih velikih jezičnih modela.

Ovaj rad fokusira se na usporedbu performansi različitih velikih jezičnih modela kroz primjenu u alatu Promptly. Kroz analizu i evaluaciju ovih modela, cilj je istražiti njihovu sposobnost u različitim zadacima obrade prirodnog jezika te identificirati prednosti i ograničenja svakog modela. Osim toga, istražuje se i kako različite arhitekture i konfiguracije modela i njihove performanse u kontekstu Promptly alata.

U ovom uvodu isticanje se značaja obrade prirodnog jezika u suvremenom društvu, naglašava važnost velikih jezičnih modela te postavlja temelj za analizu performansi kroz primjenu u alatu Promptly.

1. Razumijevanje domene

Autori Iannantuono i kolege, (2024) iskazuju da veliki jezični modeli (LLM-ovi) predstavljaju nedavni napredak u generativnoj umjetnoj inteligenciji, omogućujući računalima interpretaciju i generiranje ljudski sličnog teksta. Njihova sposobnost stvaranja novog sadržaja potaknula je istraživanje primjena u medicini i zdravstvu, posebno u pružanju informacija o karcinomu. Iako se smatraju korisnim "virtualnim asistentima", postoji značajna razina pogrešaka koja ih prati. U kontekstu imunoterapije u onkologiji, LLM-ovi se smatraju potencijalno korisnim alatima za pružanje informacija pacijentima i zdravstvenim stručnjacima. Stoga su prikazane sposobnosti nekoliko LLM-ova za pružanje edukativnih i upravljačkih informacija u ovom području.

Usporedba performansi različitih velikih jezičnih modela postala je važna tema istraživanja i razvoja u kontekstu alata kao što je Promptly. Promptly, kao sučelje koje omogućuje korištenje LLM-ova za rješavanje različitih zadataka putem upita, zahtijeva detaljnu analizu kako bi se odabrao najprikladniji model za specifične potrebe korisnika.

U istraživanju autor Zhao i kolege, (2023), pregledan je nedavni napredak velikih jezičnih modela (LLM-ova), a obrađeni su ključni koncepti, nalazi i tehnike za razumijevanje i korištenje LLM-ova. Fokus je bio na modelima velikih veličina, s naglaskom na izazove u pre-treniranju, prilagodbi, korištenju i evaluaciji LLM-ova. Raspravljene su metode i resursi za razvoj LLM-ova, kao i smjernice za njihovu implementaciju.

* Osnove i principi: LLM-ovi uče iz velikih tekstualnih podataka putem nenadziranog pre-treniranja. No, kako jednostavni ciljevi modeliranja jezika omogućuju LLM-ovima rješavanje raznih stvarnih zadataka i dalje je izazov.
* Arhitektura modela: Transformer je standardna arhitektura za LLM-ove, ali poboljšanje njegove efikasnosti i performansi predstavlja izazov.
* Treniranje modela: Za pre-treniranje LLM-ova ključno je uspostaviti infrastrukturu i postupak treniranja koji podržava sustavno prikupljanje, čišćenje i organizaciju podataka. Potrebno je razviti ekonomične pristupe za optimizaciju LLM-ova i dijeliti više principa i recepta za trening kako bi se smanjio rizik od degradacije ili neuspjeha.
* Korištenje modela: Korištenje LLM-ova putem prirodno-jezičnih upita postalo je glavni pristup rješavanju različitih zadataka. Potrebno je istražiti kako pravilno postaviti upite kako bi se dobio željeni odgovor od LLM-ova te smanjiti troškove inferencije.
* Sigurnost i usklađenost: LLM-ovi se suočavaju s izazovima sigurnosti u praktičnoj primjeni. Metode usklađenosti su ključne za sprječavanje problema, ali je potrebno poboljšati njihovu učinkovitost i pristupnost te razmotriti pitanja privatnosti.
* Primjena i ekosustav: LLM-ovi pokazuju sposobnosti u rješavanju različitih zadataka i mogu imati značajan utjecaj na pristup informacijama i tehnološke inovacije. Ipak, sigurnost mora biti primarni prioritet u razvoju ovih sustava za pozitivan utjecaj na društvo.

Sve navedeno su korisne informacije koje je potrebno imati u vidu prije rada u razvijanju i radu sa LLM-ovima.

Različiti izbori arhitekture i zadataka pre-treniranja mogu rezultirati različitim sposobnostima i sklonostima kod jezičnih modela. Autor Zhao i kolege, (2023) navode da se većina trenutnih jezičnih modela temelji na uzročnoj dekoder arhitekturi, no nedostaje detaljna analiza njenih prednosti u odnosu na druge opcije. Postojeći radovi pokazuju da uzročna dekoder arhitektura može imati izvanredne sposobnosti generalizacije na primjerima koji nisu bili dio treninga. Također, otkriven je zakon skaliranja kod uzročnih dekodera, što znači da se poboljšanje performansi može postići povećanjem veličine modela, skupa podataka i računalnih resursa. Potrebno je više istraživanja kako bi se bolje razumjelo kako izbor arhitekture i pre-treniranih zadataka utječu na sposobnosti jezičnih modela, posebno kod enkoder-dekoder arhitektura. Preporučuju raznoliko istraživanje u dizajnu arhitekture, uz istraživanje detaljnih konfiguracija jezičnih modela.

Opisane su tri glavne arhitekture velikih jezičnih modela (LLM-ova): enkoder-dekoder, uzročni dekoder i prefiksni dekoder. Enkoder-dekoder arhitektura, koja se temelji na Transformer modelu, koristi se za različite zadatke obrade prirodnog jezika. Uzročna dekoder arhitektura, koja je karakteristična za GPT-seriju modela, koristi se za autoregresivno generiranje teksta. Prefiksna dekoder arhitektura, koju koriste modeli poput GLM130B i U-PaLM, omogućuje bidirekcionalnu pažnju nad prefiksnim tokenima i unidirekcioniranu pažnju samo na generirane tokene. Nadalje, istražuju se novi pristupi poput mješavine stručnjaka skaliranja i nastalih arhitektura, koje poboljšavaju efikasnost modeliranja jezika i omogućuju paralelno kodiranje cijele rečenice. Ovi novi pristupi imaju kapacitet enkodiranja cijele rečenice paralelno, a istovremeno omogućuju rekurzivno generiranje izlaza, čime se poboljšava efikasnost procesa dekodiranja.

U dijelu "A Comprehensive Overview of Large Language Models" od strane Naveed i kolega, (2023) pružena je također opsežna analiza komponenti arhitekture i strategija obuke velikih jezičnih modela (LLM). Ključni nalazi uključuju značajan utjecaj manjih promjena na performanse i stabilnost, poput korištenja normalizacije slojeva i kodiranja pozicija. Također, detaljno su razmatrane strategije obuke poput mješovite preciznosti, tehnika za ublažavanje nestabilnosti tijekom obuke i metode inicijalizacije težina. Neke od ključnih točaka koje se ističu u vezi arhitekture su:

* Normalizacija slojeva: Pokazano je da normalizacija slojeva ima značajan utjecaj na performanse i stabilnost LLM-ova. Različite strategije normalizacije, poput pre-normiranja i post-normiranja, koriste se kako bi se postigla stabilnost u obuci.
* Kodiranje pozicija: Utječe na performanse i stabilnost LLM-ova, ali postoji nedoumica u literaturi o tome koja je metoda kodiranja pozicija najučinkovitija.
* Paralelna pažnja: Korištenje paralelne pažnje, gdje su pažnja i slojevi za prosljeđivanje podataka paralelni umjesto sekvencijalni, pokazalo se kao metoda koja smanjuje vrijeme obuke za 15% bez opaženog smanjenja performansi.
* Mješavina stručnjaka: Arhitektura mješavine stručnjaka omogućuje skaliranje modela na trilijun parametara i pokazuje bolje performanse od gusto povezanih modela uz manje računalnih troškova.
* Rijetki naspram gusto aktivirani modeli: Rijetki moduli poput rijetkih transformatora koriste se za smanjenje računalnih troškova, dok gusto aktivirani modeli poput onih koji koriste arhitekturu mješavine stručnjaka pokazuju bolje performanse za istu količinu podataka.

Ukratko, različite arhitektonske strategije i moduli koriste se kako bi se poboljšale performanse, smanjilo vrijeme obuke i povećala stabilnost LLM-ova.

Prethodno opisane arhitekture velikih jezičnih modela (LLM-ova) predstavljaju različite pristupe u projektiranju modela koji su osmišljeni za različite zadatke u obradi prirodnog jezika. Enkoder-dekoder arhitektura, uzročni dekoder arhitektura i prefiksna dekoder arhitektura imaju svoje karakteristike i primjene, ali sve se oslanjaju na osnovni Transformer model. Novi pristupi, poput mješavine stručnjaka (MoE) skaliranja i nastalih arhitektura, nadograđuju osnovne arhitekture kako bi poboljšali efikasnost modeliranja jezika i omogućili paralelno kodiranje cijele rečenice. Iako su ti pristupi inovativni, i dalje se temelje na prethodno razvijenim arhitekturama, što pokazuje kontinuirani razvoj i unaprjeđenje LLM-ova.

Kasneci i autori, (2023) analiziraju ključne izazove i rizike povezane s primjenom velikih jezičnih modela u obrazovanju. Razmatraju se problemi poput autorskih prava, pristranosti, prevelike ovisnosti učenika i nastavnika o modelima, nedostatka razumijevanja i ekspertize, sigurnosti podataka te održive upotrebe. Za svaki identificirani rizik predlažu se odgovarajuće strategije za njegovo suzbijanje i prevenciju, uključujući traženje dozvola za korištenje sadržaja, osiguravanje raznolikosti u podacima za treniranje, poticanje kritičkog razmišljanja, osiguranje sigurnosti podataka, smanjenje troškova održavanja modela te osiguranje pristupačnosti i pravednosti. Također se raspravlja o važnosti pravilnog dizajna korisničkih sučelja te potrebi za istraživanjem u ovom području kako bi se osiguralo da tehnologija bude primjenjiva i korisna za različite dobne skupine i potrebe korisnika. Ističe se važnost daljnjeg istraživanja i razvoja kako bi se maksimizirale koristi, minimizirali rizici te osigurala pouzdana i pravedna primjena velikih jezičnih modela u obrazovanju.

Veliki jezični modeli koje se planira koristiti za usporedbu u Promptly-u su ChatGPT i Google Gemini.

Ellen Glover, (2024) navodi da je Gemini najnoviji AI model tvrtke Google koji može prepoznati uzorke u podacima i generirati novi sadržaj na temelju tih uzoraka. Treniran je na velikom skupu podataka. Nakon obuke, koristi se Transformer arhitekturom neuronske mreže za generiranje riječi ili fraza koje slijede ulazni tekst. Također može obraditi slike, videozapise i audio. Google tvrdi da Gemini može "besprijekorno razumjeti i zaključivati" različite ulaze poput teksta na fotografiji znaka ili generiranja priče na temelju ilustracije.

Prema Mark Wilson, (2024) glavni zadatak ChatGPT-a je predviđanje sljedeće riječi u rečenici, što se temelji na učestalim obrascima u velikim količinama tekstualnih podataka na kojima je obučen. Na neki način, možemo ga opisati kao vrhunski napredni alat za ispravak teksta. Ovaj model je također obučen na ogromnim količinama tekstualnih podataka kako bi predvidio sljedeću riječ u nizu. Konačno, "transformer" je arhitektura neuronske mreže na kojoj se temelji ChatGPT, a koju su razvili istraživači iz Googlea 2017. godine. Ova arhitektura je posebno prikladna za obradu prirodnog jezika, poput odgovaranja na pitanja ili generiranja teksta. Google je istaknuo svoju ulogu u razvoju ove tehnologije, ali je ChatGPT taj koji je popularizirao koncept.

Autor Harry Guinness, (2024) navodi sljedeće modele kao najbolje LLM modele:

| **LLM** | **Developer** | **Popular apps that use it** | **# of parameters** | **Access** |
| --- | --- | --- | --- | --- |
| [**GPT**](https://zapier.com/blog/best-llm/#gpt) | OpenAI | Microsoft, Duolingo, Stripe, Zapier, Dropbox, ChatGPT | 175 billion+ | API |
| [**Gemini**](https://zapier.com/blog/best-llm/#gemini) | Google | Some queries on Bard | Nano: 1.8 & 3.25 billion; others unknown | API |
| [**PaLM 2**](https://zapier.com/blog/best-llm/#palm) | Google | Google Bard, Docs, Gmail, and other Google apps | 340 billion | API |
| [**Llama 2**](https://zapier.com/blog/best-llm/#llama) | Meta | Undisclosed | 7, 13, and 70 billion | Open source |
| [**Vicuna**](https://zapier.com/blog/best-llm/#vicuna) | LMSYS Org | Chatbot Arena | 7, 13, and 33 billion | Open source |
| [**Claude 2**](https://zapier.com/blog/best-llm/#claude) | Anthropic | Slack, Notion, Zoom | Unknown | API |
| [**Stable Beluga**](https://zapier.com/blog/best-llm/#stable-beluga) | Stability AI | Undisclosed | 7, 13, and 70 billion | Open source |
| [**StableLM**](https://zapier.com/blog/best-llm/#stablelm) | Stability AI | Undisclosed | 7, 13, and 70 billion | Open source |
| [**Coral**](https://zapier.com/blog/best-llm/#coral) | Cohere | HyperWrite, Jasper, Notion, LongShot | Unknown | API |
| [**Falcon**](https://zapier.com/blog/best-llm/#falcon) | Technology Innovation Institute | Undisclosed | 1.3, 7.5, 40, and 180 billion | Open source |
| [**MPT**](https://zapier.com/blog/best-llm/#mpt) | Mosaic | Undisclosed | 7 and 30 billion | Open source |
| [**Mixtral 8x7B**](https://zapier.com/blog/best-llm/#mixtral) | Mistral AI | Undisclosed | 46.7 billion | Open source |
| [**XGen-7B**](https://zapier.com/blog/best-llm/#xgen) | Salesforce | Undisclosed | 7 billion | Open source |
| [**Grok**](https://zapier.com/blog/best-llm/#grok) | xAI | Grok Chatbot | Unknown | Chatbot |

Table 1 Popis najboljih LLM modela (preuzeto od autora Harry Guinnessa)

U dijelu "Performance Comparison of Large Language Models on VNHSGE English Dataset: OpenAI ChatGPT, Microsoft Bing Chat, and Google Bard" od strane Dao, (2023) istražuje se performansa tri velika jezična modela (LLM-ova) - OpenAI ChatGPT, Microsoft Bing Chat i Google Bard, na VNHSGE engleskom skupu podataka. Cilj je usporediti performanse ovih modela u odgovaranju na engleski jezični test koji se koristi u visokim školama u Vijetnamu. Autor uspoređuje preciznost odgovora svakog modela i zaključuje da je Bing Chat najbolji u usporedbi s ChatGPT-om i Bardom. Također se ističe da su ovi modeli nadmašili vijetnamske studente u engleskoj jezičnoj sposobnosti.

Točniji rezultati pokazuju da je performansa modela BingChat bila najbolja, s preciznošću od 92.4%, dok su performanse modela Bard bile nešto niže, s preciznošću od 86%. OpenAI ChatGPT (GPT-3.5) je imao preciznost od 79.2%. Dakle, BingChat se istaknuo kao najbolji u ovom istraživanju, sugerirajući da bi BingChat i Bard mogli zamijeniti ChatGPT u vijetnamskim školama, s obzirom na trenutnu nedostupnost ChatGPT-a u Vijetnamu.

A chart with colorful squares and lines

Description automatically generated with medium confidence

Slika 1 Stabilnost izvedbe LLM-a (preuzeto od autora Dao X)

A graph of numbers and a number of bars

Description automatically generated with medium confidence

Slika 2 Usporedba uspješnosti LLM-a i vijetnamskih studenata (Preuzeto od autora Dao X)

1. Razumijevanje podataka

Skup podataka koji se koristi je iz područja video igara. Nadalje se objasne podaci koji se budu koristili u daljnjem radu. Projekt koristi skup podataka "Video Games Sales as at 22 Dec 2016" od ANDRESHG, (2021) kako bi se prikazalo strukturu podataka i omogućilo analizu koja će poslužiti za odgovaranje na različita pitanja. Uz korištenje LLM (Large Language Model), podaci će se koristiti za generiranje odgovora na postavljena pitanja, pružajući korisnicima korisne informacije i uvide o svijetu video igara.

Analiza ovog skupa podataka omogućit će razumijevanje ključnih aspekata industrije video igara, uključujući popularne žanrove, preferencije različitih tržišta, utjecaj ocjena kritičara i korisnika na prodaju igara te mnoge druge važne faktore. Korištenjem LLM, istraživanje će omogućiti brzu obradu i interpretaciju podataka, što će rezultirati brzim i relevantnim odgovorima na pitanja korisnika.

Ovaj pristup analizi podataka omogućuje interaktivnu komunikaciju s informacijama, pružajući korisnicima mogućnost postavljanja pitanja o video igrama i dobivanja relevantnih odgovora. Kroz ovaj proces, cilj je pružiti korisnicima obogaćeno iskustvo istraživanja i otkrivanja novih informacija o svijetu video igara.

Prikazan je skup podataka prvih par podataka.

A screenshot of a computer screen

Description automatically generated

Slika 3 Prikaz tablice podataka

A screenshot of a computer

Description automatically generated

Slika 4 Prikaz tablice podataka

Objašnjenja atributa:

Name: Naziv igre  
Platform: Platforma na kojoj je igra objavljena (npr. Wii, NES, GB)  
Year\_of\_Release: Godina izdanja igre  
Genre: Žanr igre (npr. Sports, Platform, Racing, Role-Playing, Puzzle)  
Publisher: Izdavač igre  
NA\_Sales: Prodaja u Sjevernoj Americi (u milijunima primjeraka)  
EU\_Sales: Prodaja u Europi (u milijunima primjeraka)  
JP\_Sales: Prodaja u Japanu (u milijunima primjeraka)  
Other\_Sales: Prodaja u ostalim dijelovima svijeta (u milijunima primjeraka)  
Global\_Sales: Ukupna svjetska prodaja igre (u milijunima primjeraka)  
Critic\_Score: Ocjena kritičara (ako je dostupna)  
Critic\_Count: Broj ocjena kritičara (ako je dostupno)  
User\_Score: Ocjena korisnika (ako je dostupna)  
User\_Count: Broj korisnika (ako je dostupno)  
Developer: Razvojni tim koji je kreirao igru  
Rating: ESRB rejting igre (npr. E - Everyone, T - Teen, M - Mature)

Opis atributa:

A screenshot of a computer

Description automatically generated

Slika 5 Prikaz atributa (iz google colab-a)

A screenshot of a graph

Description automatically generated

Slika 6 Prikaz statistike brojčani atributa (iz google colab-a)

Objašnjenje:

Year\_of\_Release:

* Prosječna godina izdanja igara u skupu podataka je 2006.49, s standardnom devijacijom od 5.88 godina. To znači da su većina igara izdana između 2003. i 2010. godine, dok su najstarije igre iz 1980. godine, a najnovije iz 2020.

Prodaja po regijama (NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, Global\_Sales):

* Prosječna prodaja igara u Sjevernoj Americi je 0.26 milijuna primjeraka, u Europi 0.15 milijuna primjeraka, u Japanu 0.08 milijuna primjeraka, dok je prosječna prodaja u ostalim dijelovima svijeta 0.05 milijuna primjeraka. Ukupna svjetska prodaja igara je prosječno 0.53 milijuna primjeraka.

Ocjene kritičara (Critic\_Score):

* Prosječna ocjena kritičara je 68.97, s standardnom devijacijom od 13.94. Ovo sugerira da su igre u našem datasetu generalno dobro ocijenjene, s najnižom ocjenom od 13 i najvišom ocjenom od 98.

Broj ocjena kritičara (Critic\_Count):

* Prosječan broj ocjena kritičara po igri je 26.36, s standardnom devijacijom od 18.98. To znači da većina igara ima između 12 i 36 ocjena kritičara.

Ocjene korisnika (User\_Score):

* Prosječna ocjena korisnika je 7.13, s standardnom devijacijom od 1.50. Najniža ocjena je 0, dok je najviša ocjena 9.70.

Broj korisnika (User\_Count):

* Prosječan broj ocjena korisnika po igri je 162.23, s standardnom devijacijom od 561.28. To sugerira veliku varijabilnost u broju ocjena između igara, s minimalnim brojem od 4 ocjene i maksimalnim brojem od 10665 ocjena

Nakon ovog slijedi grafički prikaz podataka:

Histogrami - nadolazećim histogramima možemo vidjeti koliko često su video igrice prodavane u određenim rasponima, odnosno recimo koliko je igrica prodano u rasponu od 2 do 2.2 milijuna.

A graph of a number of blue bars

Description automatically generated with medium confidence

Slika 7 Histogram Year\_of\_Release

.

Ovaj graf je malo drugačiji od drugih narednih. Na njemu se može vidjeti koliko često su igrice objavljene u nekim godinama. Može se vidjeti da se smatra unimodalanim ali se može interpretirati kao multimodalan, te je iskrivljen u lijevo.

A graph of a graph

Description automatically generated with medium confidence

Slika 8 Histogram NA\_Sales

Grafikon prikazuje distribuciju prodaje igara u Sjevernoj Americi. Distribucija prodaje je unimodalna, što znači da postoji jedan jasan vrhunac u distribuciji. Primjećuje se da je distribucija iskrivljena u desno. Značajno je primijetiti "stršilo" s desne strane grafikona, što ukazuje na prisutnost nekoliko iznimno popularnih igara s vrlo visokim brojem prodanih primjeraka. Ova analiza pruža dublji uvid u tržište igara u Sjevernoj Americi.

A graph with numbers and a bar chart

Description automatically generated with medium confidence

Slika 9 Histogram EU\_Sales

Grafikon prikazuje distribuciju prodaje igara u Europi. Primjećuje se da je distribucija prodaje unimodalna, što znači da postoji jedan jasan vrhunac u distribuciji. Također, grafikon je iskrivljen u desno sto ukazuje da je prodaja za manji broj igara izuzetno visoka. Značajno je primijetiti "stršilo" s desne strane grafikona, što ukazuje na prisutnost nekoliko iznimno popularnih igara s vrlo visokim brojem prodanih primjeraka. Ova analiza pruža dublji uvid u ponašanje potrošača na europskom tržištu igara.

A graph of a bar graph

Description automatically generated with medium confidence

Slika 10 Histogram JP\_Sales

Grafikon prikazuje distribuciju prodaje igara u Japanu. Primjećuje se da je distribucija prodaje unimodalna, što znači da postoji jedan jasan vrhunac u distribuciji. Osim toga, grafikon je iskrivljen u desno, što ukazuje na to da postoji veći broj igara s manjim brojem prodaja u Japanu, dok su prodaje za manji broj igara izuzetno visoke. Ova analiza može biti korisna za razumijevanje ponašanja potrošača na japanskom tržištu igara te za planiranje marketinških strategija i distribucije igara u skladu s tim trendovima ili u našem slučaju razumijevanju podataka za preporuku igara ili odgovor na sličan tip pitanja.

Horizontalni trakasti grafovi - ovaj tip grafa prikazuje distribuciju veličina grupa temeljenih na kategorijama ili faktorima.

A graph with text on it

Description automatically generated with medium confidence

Slika 11 Horizontalni trakasti grafikon Publisher

Grafikon prikazuje distribuciju broja igara prema različitim izdavačima. Na grafikonu su prikazani pet glavnih izdavača: Nintendo, Activision, Take-Two Interactive, Sony Computer Entertainment i Microsoft Game Studios. Nintendo značajno dominira s najvećim brojem izdanih igara, što sugerira da Nintendo ima jaku prisutnost na tržištu igara. Slijede Activision i Take-Two Interactive, pri čemu Activision ima nešto manji broj igara u usporedbi s Nintendo, dok Take-Two Interactive ima još manje igara u usporedbi s Activisionom. Sony Computer Entertainment i Microsoft Game Studios su manje zastupljeni, s najmanjim brojem izdanih igara u odnosu na ostale izdavače. Ovaj graf pruža uvid u relativne udjele različitih izdavača na tržištu igara te može biti koristan za analizu konkurentske pozicije i strategija na tržištu igara.

A graph with different colored bars

Description automatically generated with medium confidence

Slika 12 Horizontalni trakasti grafikon Rating

Grafikon prikazuje distribuciju ocjena igara prema klasifikaciji 'Rating'. Ocjene su podijeljene u nekoliko kategorija, uključujući 'E' (Everyone), 'T' (Teen), 'M' (Mature). Najveći broj igara ocijenjen je kao prikladan za sve uzraste ('E'), što sugerira da većina igara cilja širu publiku. Ocjenjivanje za odrasle ('M') također ima značajan udio, dok je broj igara ocijenjenih kao prikladnih samo za tinejdžere ('T') znatno manji. Ovo može biti korisna spoznaja za donositelje odluka u industriji igara prilikom planiranja ciljnih publika i marketinških strategija.

A graph of a graph

Description automatically generated

Slika 13 Scatter plot matrix numeričkih vrijednosti

Na scatter plot matrix-u, neki scatter plotovi pokazuju jasne povezanosti između varijabli, dok drugi ne pokazuju očite uzorke. Povezanost se može očitovati u obliku jasnih trendova ili uzoraka, poput linearnih ili nelinearnih veza među varijablama. Suprotno tome, na nekim scatter plotovima može se primijetiti slučajno raspršenje točaka, što sugerira nepostojanje jasne povezanosti između varijabli.

Ovi uočeni odnosi na scatter plot matrix-u pružaju početni uvid u potencijalne veze između varijabli u datasetu. Daljnja analiza i istraživanje ovih veza može zahtijevati korištenje dodatnih statističkih metoda i modela radi boljeg razumijevanja prirode odnosa između varijabli.

Kasnije radu koristi se manji skup podataka radi ograničenja koje ima promptly te google gemini tijekom tokenizacije. Uzeli smo oko prilike prvih 1000 podataka kako bi prikazali i usporedili rad različitih LLM modela.

1. Pregled Alata

Promptly je platforma bez koda koja omogućava korisnicima jednostavno stvaranje generativnih AI aplikacija i chatbotova, a podržava sve glavne pružatelje modela, uključujući:

• OpenAI

• Cohere

• Stability AI

• Hugging Face

• druge

Promptly omogućuje povezivanje više modela kako bi se kreirale vrlo dobre generativne aplikacije, uključujući modele različitih pružatelja.

Korisnici mogu uvoziti vlastite podatke iz raznih izvora kako bi se povezali s LLM modelima i stvorili napredne generativne AI aplikacije i chatbotove. Ti izvori mogu biti:

• web URL-ova

• mapa web stranica

• PDF-ova

• audio datoteka

• PPT-ova

• Drugi

Glavni atribut koji izdvaja Promptly je mogućnost jednostavne integracije njegovih prilagodljivih chat widgeta na vašu web stranicu. Ovi widgeti omogućuju korisnicima da brzo razviju konverzacijske AI aplikacije ili dodaju chatbote direktno na svoje online platforme.

Promptly također nudi opciju playground za eksperimentiranje s različitim modelima. Korisnici mogu isprobati chatbotove drugih kreatora putem marketplace-a. Tijekom izrade, dostupni su predlošci za različite aplikacije, uključujući: prevoditelja jezika, chat u stvarnom vremenu s avatarom, chatbot za web stranicu, chat s datotekama, glasovni chat, chatbot s pristupom internetu, generator marketinškog sadržaja, glasovne odgovore, AI SDR, ekstraktor podataka, chat s likovima, te sažimanje glasa. Dodatno, dostupni su prazni predlošci za: web aplikaciju, chatbota i agenta.

Glavni cilj Promptly platforme je olakšati izradu AI aplikacija i chatbotova bez potrebe za programiranjem, nudeći istovremeno snažne i prilagodljive opcije za razvoj i prilagodbu ovih aplikacija.

Iako Promptly pruža relativno jednostavan korisnički doživljaj, postoje određeni izazovi koji ga prate. Na primjer, primijetilo se probleme s promjenom boje poruka unutar chata, gdje odabrana boja ponekad ne bude primijenjena kako je očekivano. Također, iako postoji opcija za uređivanje CSS-a, također se primijetilo da implementacija CSS-a nije moguća, unatoč ponuđenoj opciji. Ovo može dovesti do zbunjenosti korisnika, s obzirom na neusklađenost između ponuđenih opcija i njihove funkcionalnosti. Unatoč tim izazovima, izrada osnovnog chatbota unutar Promptly platforme nije previše složena.

1. Prikaz strukture chatbot modela

U nastavku će biti prikazana struktura chatbot modela i od OpenAI modela i Google GEMINI modela iako je slična bude prikazana pomoću slike i mogućnosti vlastite implementacije. Budu prikazane sve unesene stavke i objašnjenja. Sama struktura je jednaka osim dijela gdje se dodaje LLM model tako da se kod jednog modela prikaže cijela struktura dok kod drugog LLM bude prikazano samo dio koji je različit.

* 1. Google Gemini

A screenshot of a computer

Description automatically generated

Slika 14 Prikaz strukture chatbot modela GEMINI

A screenshot of a computer

Description automatically generated

Slika 15 Prikaz strukture chatbot modela GEMINI

U prikazu je vidljiv 2k2\_merged.pdf i to je uređeni pdf kao što je navedeno te je dodan još jedan vlastoručni opis razvojnih timova tako da bih imali više podataka različitih za pitat (pdf je dostupan uz ovaj dokument). Taj pdf koristi kao bazu podataka tj. znanje koje izvlači i na osnovu njih daje korektne odgovore.

A screenshot of a chat

Description automatically generated

Slika 16 Prikaz strukture chatbot modela GEMINI

Prikazan je poruka sistema tj, zadatak koji dobije LLM model tijekom interakcije s korisnikom. Upute trebaju bit detaljne i korektne kako bih imali što bolje rezultate. Tijekom testiranja oba modela će imati jednake upute kao što bude primjetno.

A screenshot of a computer

Description automatically generated

Slika 17 Prikaz strukture chatbot modela GEMINI

Kod alata Gemini potrebno je unijeti dodatne opcije kao što je safety settings što je korisna opcija tako da osiguramo sigurno mjesto za razgovor sa chatbotom.

A screenshot of a computer

Description automatically generated

Slika 18 Prikaz strukture chatbot modela GEMINI

* 1. OpenAI chatbot

Kod OpenAI chatbot ima jednaku strukturu jedina razlika je samo korištenje LLM modela te radi jednostavnosti i preglednosti dokumenta bude prikazano samo različiti dijelovi u modelu.

A screenshot of a computer

Description automatically generated

Slika 19 Prikaz strukture chatbot modela OpenAI

A screenshot of a computer

Description automatically generated

Slika 20 Prikaz strukture chatbot modela OpenAI

1. Prikaz rada chatbota

U nastavku će biti prikazane slike razgovora s različitim LLM modelima tj. OpenAI i Google GEMINI modelima. Princip rada je da postavimo jednaka pitanja za oba model nad jednakim podacima. Nakon toga pokušamo testirati odgovore i povijest pamćenja razgovora, a u sljedećem poglavlju tj. usporedba performansi će konkretno navedene prednosti i mane te primijećeni rezultati.

* 1. Razgovor s Google GEMINI

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a chat

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a chat

Description automatically generated

* 1. Razgovor s OpenAI

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

1. Usporedba Performansi

Tijekom testiranja modela i razgovora sa chatbotom bilježili su se performanse te će se u nastavku navesti rezultati na osnovu kojih će se moći donijeti koristan zaključak.

Testiranje i analiza dva različita LLM modela provedena su korištenjem Promptly alata. U oba slučaja, modeli su bili postavljeni na zadatke preporuke iz skupa podataka, procjene kvalitete rezultata, identificiranja halucinacija te praćenja povijesti razgovora. Svi testovi su provedeni na hrvatskom jeziku. Rezultati i zapažanja su sljedeći:

**Google GEMINI**

Preporuke i Rezultati:

* Model je uspješno odgovarao na jednostavna pitanja, no često je dodavao opis igre iako to nije bilo zatraženo.
* Promptly alat za ovaj model nije imao opciju podešavanja kreativnosti, što je ograničavalo njegove performanse.

Halucinacije i praćenje povijesti:

* Model je pokazao slabosti u praćenju povijesti razgovora. Na jednostavna pitanja koja su se mogla odgovoriti iz prethodnih poruka, odgovarao je dodatnim pitanjima, ponekad na ćirilici.
* U slučaju kada nije znao odgovor, model bi sam sebi postavljao pitanja i odgovarao na njih, te prešao na ćirilicu.
* Prilikom upita o igrama s visokim ocjenama, model je davao izmišljene metrike koje nisu bile utemeljene u podacima.

Brzina Generiranja:

* Oba modela su pokazala jednaku brzinu generiranja odgovora pod normalnim okolnostima.
* Međutim, kada je skup podataka bio dvostruko veći, oba modela su postala vrlo spora, osobito kod prvog pitanja.
* Prvi LLM model (Google GEMINI) je imao tendenciju "pucanja" pri duljem čekanju na učitavanje modela, što je značajno smanjilo njegovu pouzdanost.

**OpenAI model**

Preporuke i Rezultati:

* Model je također odgovarao na pitanja i dodavao opis igara, no Promptly alat za ovaj model omogućava podešavanje kreativnosti, što poboljšava njegove performanse.
* Rezultati su bili bolji, primjerice, model je prepoznao postojanje Namco studija i njegovo spajanje s Bandai Namco, što prvi model nije.

Halucinacije i praćenje povijesti:

* Kod pitanja o igrama s visokim ocjenama, tri rezultata su bila halucinirana, dok su ostali bili preuzeti iz PDF dokumenata, bez izmišljenih ocjena i metrika.
* Iako ni ovaj model nije ispunio očekivanja u praćenju povijesti razgovora, dao je korisne odgovore.

Brzina Generiranja:

* Drugi model (OpenAI) je također pokazao smanjenje brzine pri radu s većim skupom podataka, no nije imao problema s "pucanjem" kao prvi model.
* Iako sporiji pri većim podacima, drugi model je bio stabilniji i pouzdaniji.

Zaključak

Oba LLM modela su pokazala korisnost, no drugi model (OpenAI) se pokazao boljim. Mogućnost podešavanja kreativnosti ovog modela omogućila bi ga usavršiti prema našim željama i potrebama. Drugi model (OpenAI) daje smislenije odgovore i bolje prati podatke iz PDF dokumenata. Nakon promjene parametara, drugi model je davao još bolje rezultate, dok je prvi model (GEMINI) nema tu opciju i nastao bi problem posebno kada bi naišao na nepoznate upite. Također, prvi model (GEMINI) slabo raspoznaje jezike ovih prostora što može biti razlog za promjenu pisma na ćirilicu.

Što se tiče brzine generiranja, oba modela su pokazala slične performanse u standardnim uvjetima. No, pri radu s većim skupovima podataka, oba modela su postala sporija, s time da je prvi model imao ozbiljne probleme s stabilnošću, dok je drugi model ostao pouzdan.

Sve u svemu, drugi LLM model nudi veću prilagodljivost, preciznost i stabilnost u odgovaranju na pitanja, čineći ga boljim izborom za daljnju uporabu i istraživanje.

Dodatne Napomene o Korištenju Promptly Extractora i Flowisu

Tijekom rada u Promptly alatu, ponuđen je Promptly extractor podataka koji odmah ima ugrađen LLM model od OpenAI. Ovaj model pruža još bolje rezultate od prijašnjih testiranih modela i značajno je brži. Iako nije korišten u testiranju radi pravednosti i vrijednosti ispitivanja ostalih LLM modela, činjenica je da LLM od OpenAI ima više mogućnosti, naročito u samoj strukturi modela. Promptly-ev extractor s ugrađenim LLM modelom od OpenAI bio bi još bolje rješenje za naše potrebe.

Također, na drugom kolegiju pod nazivom "Inteligentni interaktivni sustavi", razvijamo chatbot u alatu Flowise. Tamo smo radili na istu temu videoigara s istim skupom podataka i istim LLM modelom od OpenAI. Rezultati su bili daleko bolji od onih dobivenih s prethodnim modelima, a jednostavnost korištenja je, po našem mišljenju, praktički jednaka, ako ne i bolja.

Sve u svemu, korištenje LLM modela od OpenAI u kombinaciji s alatima poput Promptly extractora i Flowise čini se kao superiorno rješenje za razvoj inteligentnih sustava temeljenih na umjetnoj inteligenciji.

1. Zaključak

U ovom radu prikazali smo pregled trenutnih istraživanja o LLM (Large Language Models) modelima, uključujući objašnjenja što su LLM modeli, kako rade i koje su njihove primjene. Također smo odlučili provesti samostalno testiranje dva različita LLM modela kako bismo procijenili njihove performanse u preporukama iz skupa podataka, identificiranju halucinacija te praćenju povijesti razgovora.

Na temelju provedenih testiranja, prvi model (GEMINI) pokazao je određene slabosti, uključujući nepotrebno dodavanje informacija, haluciniranje odgovora, promjenu pisma na ćirilicu i loše praćenje povijesti razgovora. Drugi model (OpenAI) se pokazao boljim, pružajući preciznije odgovore, bolje praćenje podataka i veću stabilnost prilikom rada s većim skupovima podataka, a mogućnost podešavanja kreativnosti dodatno je poboljšala njegove performanse. Osim toga, mogućnost korištenje Promptly extractora s ugrađenim LLM modelom od OpenAI pokazalo je superiorne rezultate u usporedbi s oba testirana modela, pružajući brže i preciznije odgovore te značajan napredak. Na drugom kolegiju, "Inteligentni interaktivni sustavi", korištenje OpenAI modela unutar Flowise alata također je pokazalo izvanredne rezultate, uz jednostavnost korištenja koja je dodatno potvrdila superiornost ovog modela.

Za buduća istraživanja, bilo bi korisno istražiti i testirati druge alate koje nudi Promptly, te usporediti njihove rezultate s dosadašnjim modelima. Daljnja testiranja mogu pomoći u identificiranju dodatnih prednosti i nedostataka različitih LLM modela, te pružiti širi uvid u njihove performanse i mogućnosti. Nastavak istraživanja u ovom smjeru može doprinijeti razvoju još učinkovitijih i preciznijih inteligentnih sustava.

U zaključku, napredak LLM modela, osobito onih razvijenih od strane OpenAI, pokazuje značajna poboljšanja u performansama, točnosti i stabilnosti. Korištenje ovih modela u alatima poput Promptly i Flowise nudi najbolje rješenje za razvoj inteligentnih sustava temeljenih na umjetnoj inteligenciji. Naši rezultati ukazuju na to da su OpenAI LLM modeli trenutno na čelu tehnologije, pružajući vrhunske mogućnosti za različite primjene.

Link promptly GEMINI modela: <https://trypromptly.com/app/bac65cf9-8b13-47b2-ade6-1cd6258f13f8>

Link promptly OpenAI modela:

1. Popis literature

ANDRESHG. (2021). *Video Games EDA and Visualization*. Kaggle.

Dao, X.-Q. (2023). *Performance Comparison of Large Language Models on VNHSGE English Dataset: OpenAI ChatGPT, Microsoft Bing Chat, and Google Bard*.

Ellen Glover. (2024, March 13). *What Is Google Gemini?* Builtin. https://builtin.com/articles/google-gemini

Harry Guinness. (2024, January 30). *The best large language models (LLMs) in 2024*. Zapier. https://zapier.com/blog/best-llm/

Iannantuono, G. M., Bracken-Clarke, D., Karzai, F., Choo-Wosoba, H., Gulley, J. L., & Floudas, C. S. (2024). Comparison of Large Language Models in Answering Immuno-Oncology Questions: A Cross-Sectional Study. *The Oncologist*. https://doi.org/10.1093/oncolo/oyae009

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. In *Learning and Individual Differences* (Vol. 103). https://doi.org/10.1016/j.lindif.2023.102274

Mark Wilson. (2024, February 25). *ChatGPT explained – everything you need to know about the AI chatbot*. Techrada. https://www.techradar.com/news/chatgpt-explained

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023a). *A Survey of Large Language Models*.

1. Popis slika

[Slika 1 Stabilnost izvedbe LLM-a (preuzeto od autora Dao X) 8](#_Toc167984819)

[Slika 2 Usporedba uspješnosti LLM-a i vijetnamskih studenata (Preuzeto od autora Dao X) 8](#_Toc167984820)

[Slika 3 Prikaz tablice podataka 9](#_Toc167984821)

[Slika 4 Prikaz tablice podataka 9](#_Toc167984822)

[Slika 5 Prikaz atributa (iz google colab-a) 10](#_Toc167984823)

[Slika 6 Prikaz statistike brojčani atributa (iz google colab-a) 11](#_Toc167984824)

[Slika 7 Histogram Year\_of\_Release 12](#_Toc167984825)

[Slika 8 Histogram NA\_Sales 13](#_Toc167984826)

[Slika 9 Histogram EU\_Sales 14](#_Toc167984827)

[Slika 10 Histogram JP\_Sales 15](#_Toc167984828)

[Slika 11 Horizontalni trakasti grafikon Publisher 16](#_Toc167984829)

[Slika 12 Horizontalni trakasti grafikon Rating 17](#_Toc167984830)

[Slika 13 Scatter plot matrix numeričkih vrijednosti 18](#_Toc167984831)

[Slika 14 Prikaz strukture chatbot modela GEMINI 21](#_Toc167984832)

[Slika 15 Prikaz strukture chatbot modela GEMINI 21](#_Toc167984833)

[Slika 16 Prikaz strukture chatbot modela GEMINI 22](#_Toc167984834)

[Slika 17 Prikaz strukture chatbot modela GEMINI 22](#_Toc167984835)

[Slika 18 Prikaz strukture chatbot modela GEMINI 23](#_Toc167984836)

[Slika 19 Prikaz strukture chatbot modela OpenAI 23](#_Toc167984837)

[Slika 20 Prikaz strukture chatbot modela OpenAI 23](#_Toc167984838)