

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

5. Sadržaj

5. Sadržaj	iv
1. Uvod	1
2. Razumijevanje domene	2
3. Razumijevanje podataka	8
4. Pregled Alata	19
5. Prikaz strukture chatbot modela	20
5.1. Google Gemini	20
5.2. OpenAI chatbot	23
6. Prikaz rada chatbota	23
6.1. Razgovor s Google GEMINI	24
6.2. Razgovor s OpenAI	27
7. Usporedba Performansi	30
8. Zaključak	33
9. Popis literature	35
10. Popis slika	36

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.

2. 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 dekođer arhitekturi, no nedostaje detaljna analiza njenih prednosti u odnosu na druge opcije. Postojeći radovi pokazuju da uzročna dekođer arhitektura može imati izvanredne sposobnosti generalizacije na primjerima koji nisu bili dio treninga. Također, otkriven je zakon skaliranja kod uzročnih dekođera, 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-dekođer arhitektura. Preporučuju raznoliko istraživanje u dizajnu arhitekture, uz istraživanje detaljnih konfiguracija jezičnih modela.

Opisane su tri glavne arhitekture velikih jezičnih modela (LLM-ova): enkoder-dekođer, uzročni dekođer i prefiksni dekođer. Enkoder-dekođer arhitektura, koja se temelji na Transformer modelu, koristi se za različite zadatke obrade prirodnog jezika. Uzročna dekođer arhitektura, koja je karakteristična za GPT-seriju modela, koristi se za autoregresivno generiranje teksta. Prefiksna dekođer arhitektura, koju koriste modeli poput GLM130B i U-PaLM, omogućuje bidirekcionalnu pažnju nad prefiksним 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 trilion 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 dekode arhitektura i prefiksna dekode 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 unapređenje 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 "besprijeorno 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:

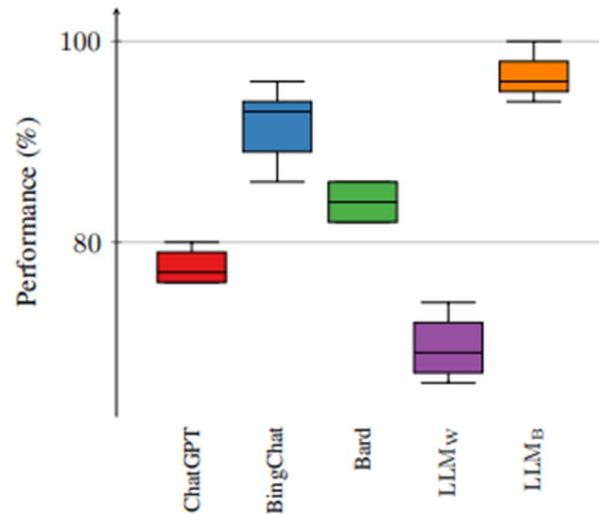
LLM	Developer	Popular apps that use it	# of parameters	Access
<u>GPT</u>	OpenAI	Microsoft, Duolingo, Stripe, Zapier, Dropbox, ChatGPT	175 billion+	API
<u>Gemini</u>	Google	Some queries on Bard	Nano: 1.8 & 3.25 billion; others unknown	API
<u>PaLM 2</u>	Google	Google Bard, Docs, Gmail, and other Google apps	340 billion	API
<u>Llama 2</u>	Meta	Undisclosed	7, 13, and 70 billion	Open source
<u>Vicuna</u>	LMSYS Org	Chatbot Arena	7, 13, and 33 billion	Open source
<u>Claude 2</u>	Anthropic	Slack, Notion, Zoom	Unknown	API
<u>Stable Beluga</u>	Stability AI	Undisclosed	7, 13, and 70 billion	Open source
<u>StableLM</u>	Stability AI	Undisclosed	7, 13, and 70 billion	Open source
<u>Coral</u>	Cohere	HyperWrite, Jasper, Notion, LongShot	Unknown	API

LLM	Developer	Popular apps that use it	# of parameters	Access
<u>Falcon</u>	Technology Innovation Institute	Undisclosed	1.3, 7.5, 40, and 180 billion	Open source
<u>MPT</u>	Mosaic	Undisclosed	7 and 30 billion	Open source
<u>Mixtral 8x7B</u>	Mistral AI	Undisclosed	46.7 billion	Open source
<u>XGen-7B</u>	Salesforce	Undisclosed	7 billion	Open source
<u>Grok</u>	xAI	Grok Chatbot	Unknown	Chatbot

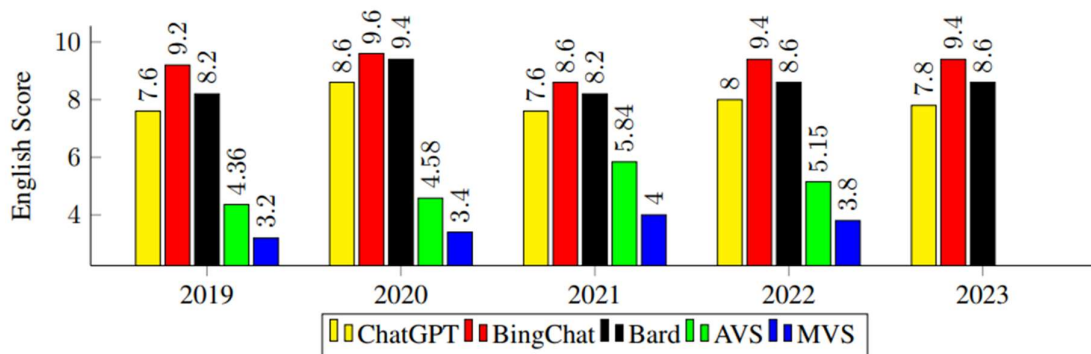
Table 1 Popis najboljih LLM modela (preuzeto od autora Harry Guinnessa)

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.



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



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

3. 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.

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN
5	Tetris	GB	1989.0	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26	NaN
6	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.28	9.14	6.50	2.88	29.80	89.0
7	Wii Play	Wii	2006.0	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92	58.0

Slika 3 Prikaz tablice podataka

Critic_Count	User_Score	User_Count	Developer	Rating
51.0	8.0	322.0	Nintendo	E
NaN	NaN	NaN	NaN	NaN
73.0	8.3	709.0	Nintendo	E
73.0	8.0	192.0	Nintendo	E
NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN
65.0	8.5	431.0	Nintendo	E

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:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16719 entries, 0 to 16718
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Name                  16717 non-null  object
1   Platform              16719 non-null  object
2   Year_of_Release       16450 non-null  float64
3   Genre                 16717 non-null  object
4   Publisher             16665 non-null  object
5   NA_Sales              16719 non-null  float64
6   EU_Sales              16719 non-null  float64
7   JP_Sales              16719 non-null  float64
8   Other_Sales           16719 non-null  float64
9   Global_Sales          16719 non-null  float64
10  Critic_Score           8137 non-null   float64
11  Critic_Count           8137 non-null   float64
12  User_Score             7590 non-null   float64
13  User_Count            7590 non-null   float64
14  Developer             10096 non-null  object
15  Rating                9950 non-null   object
dtypes: float64(10), object(6)
memory usage: 2.0+ MB
```

Slika 5 Prikaz atributa (iz google colab-a)

	count	mean	std	min	25%	50%	75%	max
Year_of_Release	16450.0	2006.487356	5.878995	1980.00	2003.00	2007.00	2010.00	2020.00
NA_Sales	16719.0	0.263330	0.813514	0.00	0.00	0.08	0.24	41.36
EU_Sales	16719.0	0.145025	0.503283	0.00	0.00	0.02	0.11	28.96
JP_Sales	16719.0	0.077602	0.308818	0.00	0.00	0.00	0.04	10.22
Other_Sales	16719.0	0.047332	0.186710	0.00	0.00	0.01	0.03	10.57
Global_Sales	16719.0	0.533543	1.547935	0.01	0.06	0.17	0.47	82.53
Critic_Score	8137.0	68.967679	13.938165	13.00	60.00	71.00	79.00	98.00
Critic_Count	8137.0	26.360821	18.980495	3.00	12.00	21.00	36.00	113.00
User_Score	7590.0	7.125046	1.500006	0.00	6.40	7.50	8.20	9.70
User_Count	7590.0	162.229908	561.282326	4.00	10.00	24.00	81.00	10665.00

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):

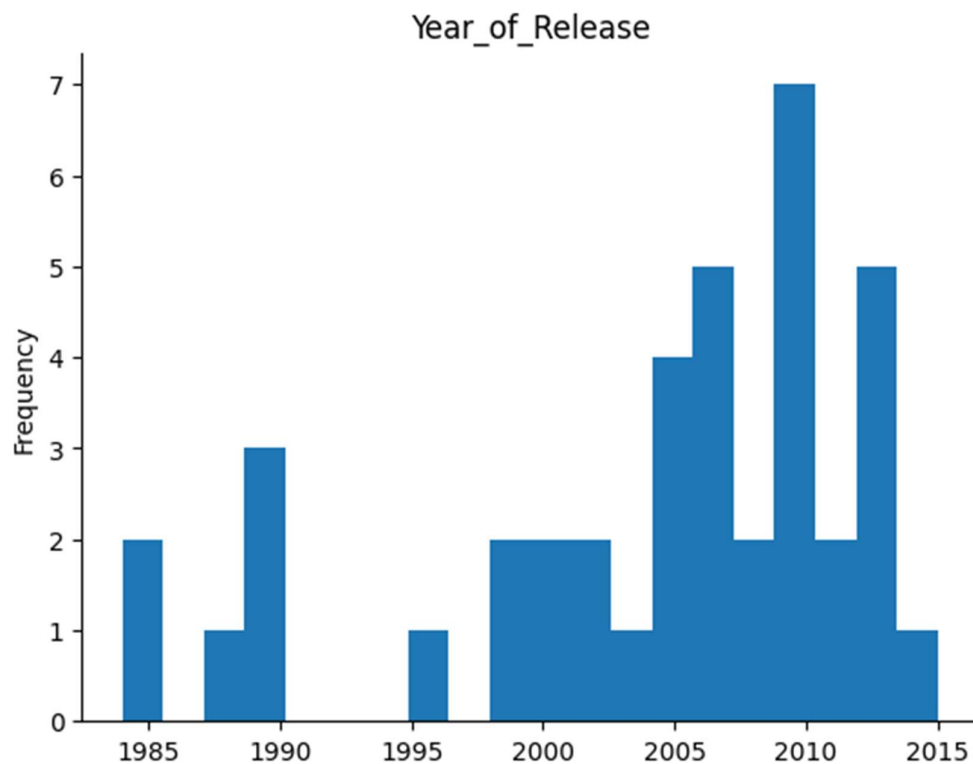
- Prosječna ocjena korisnika je 7.13, s standardnom devijacijom od 1.50. Najniža ocjena je 0, dok je najviša ocjena 9.70.

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

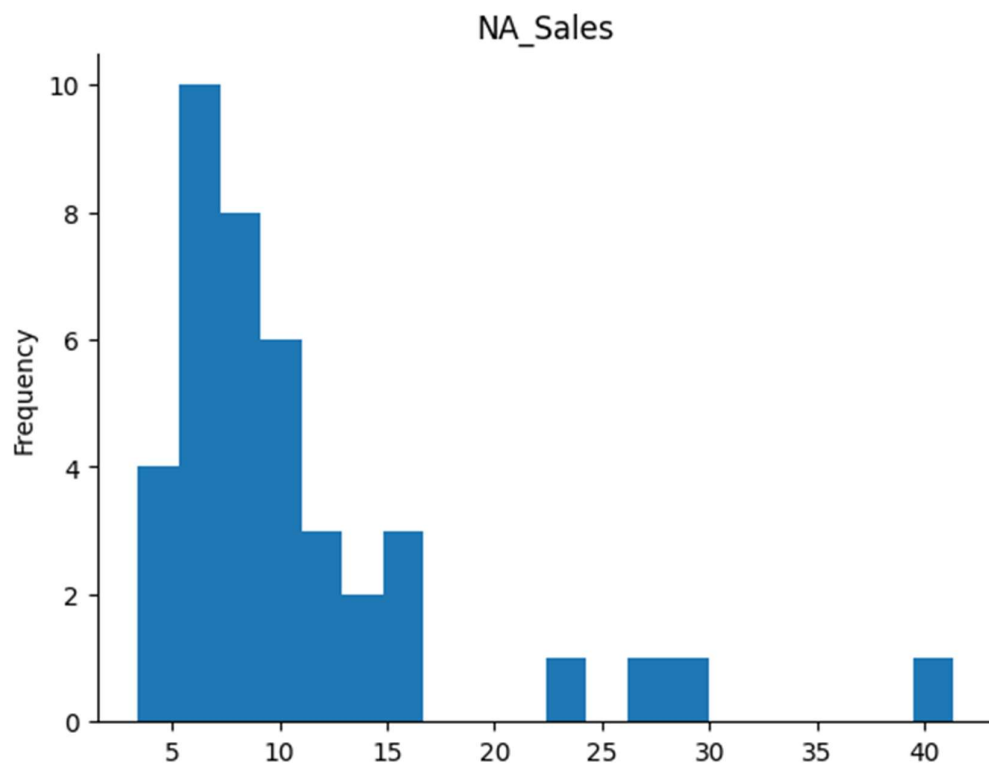
Nakon ovog slijedi grafički prikaz podataka:

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.



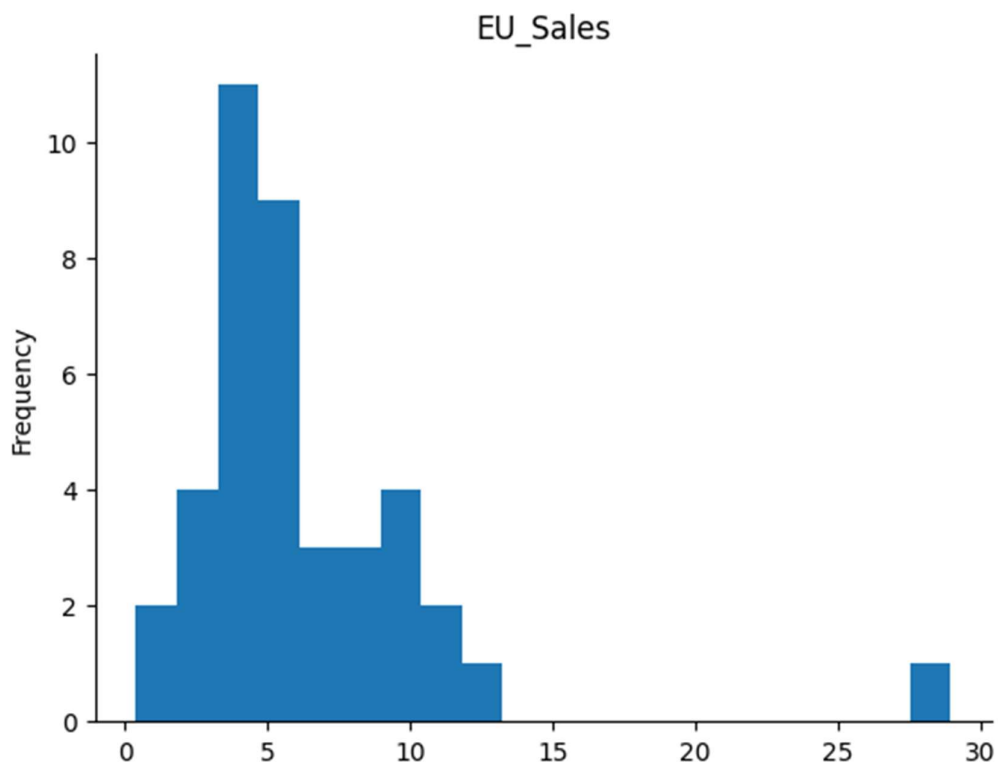
Slika 7 Histogram Year_of_Release

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 unimodalnim ali se može interpretirati kao multimodalni, te je iskrivljen u lijevo.



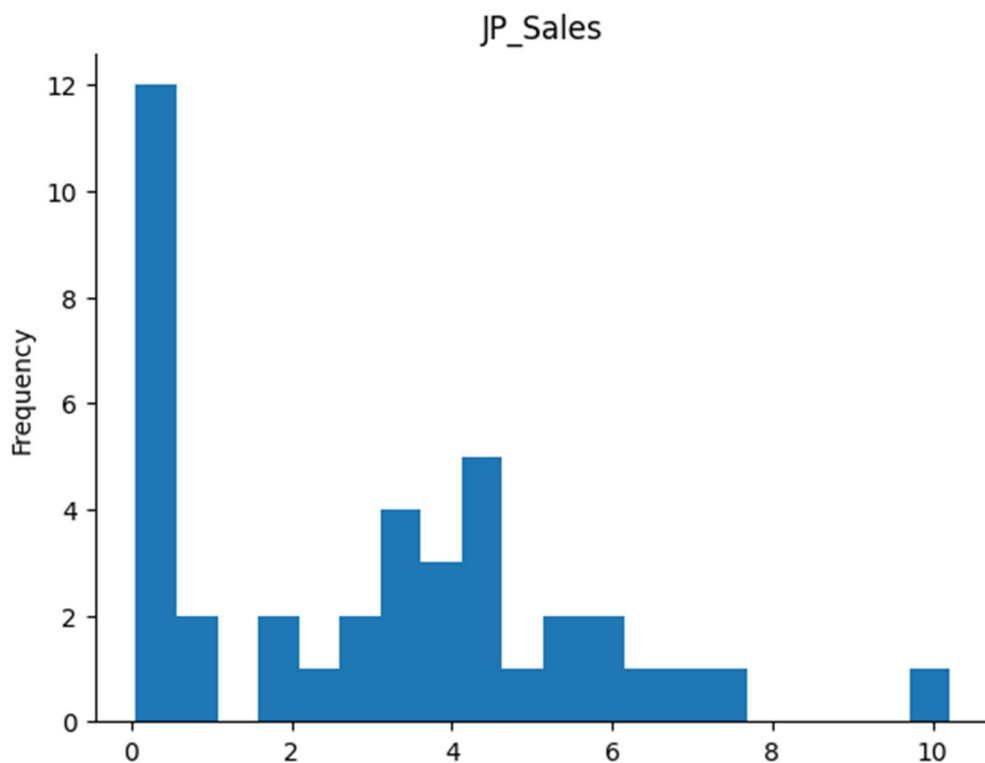
Slika 8 Histogram NA_Sales

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.



Slika 9 Histogram EU_Sales

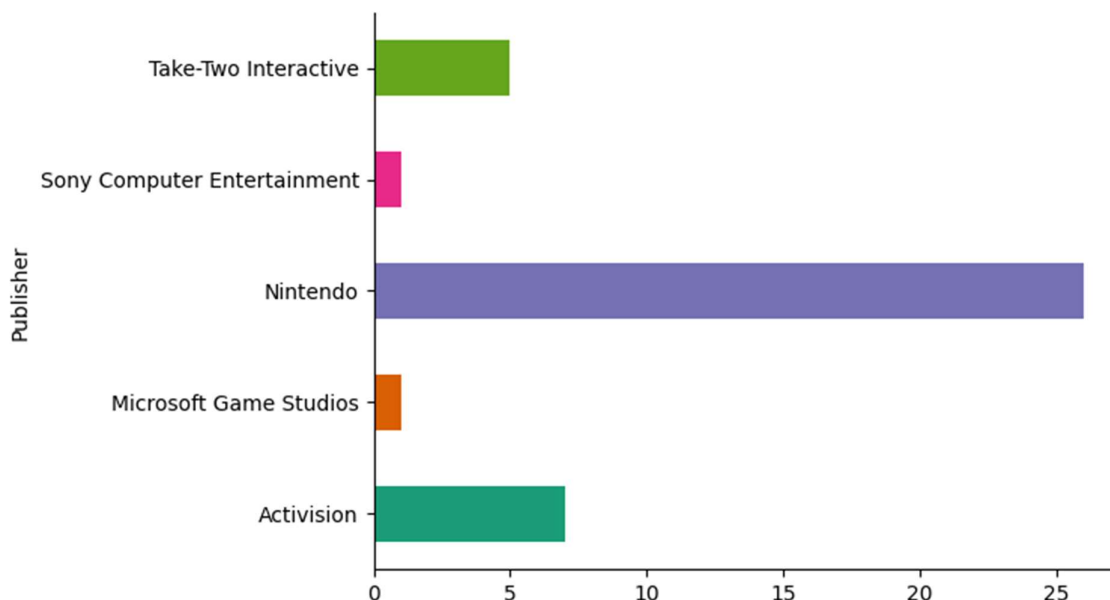
Grafikon prikazuje distribuciju prodaje igara u Europi. Primjećuje se da je distribucija prodaje unimodalna, što znači da postoji jedan jasan vrhunac u distribuciji. Također, grafikon je iskrivljen u desno što ukazuje da je prodaja za manji broj igara izuzetno visoka. Značajno je primijetiti "stršilo" 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.



Slika 10 Histogram JP_Sales

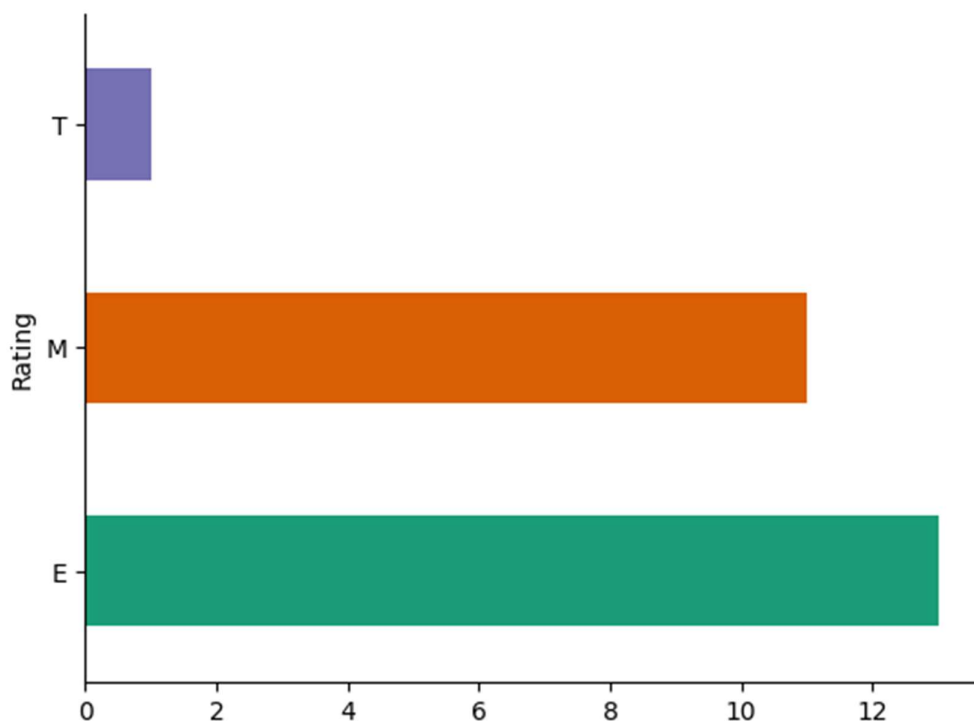
Grafikon prikazuje distribuciju prodaje igara u Japanu. Primjećuje se da je distribucija prodaje unimodalna, što znači da postoji jedan jasan vrhunac u distribuciji. Osim toga, grafikon je iskrivljen u desno, što ukazuje na to da postoji veći broj igara s manjim brojem prodaja u Japanu, dok su prodaje 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.



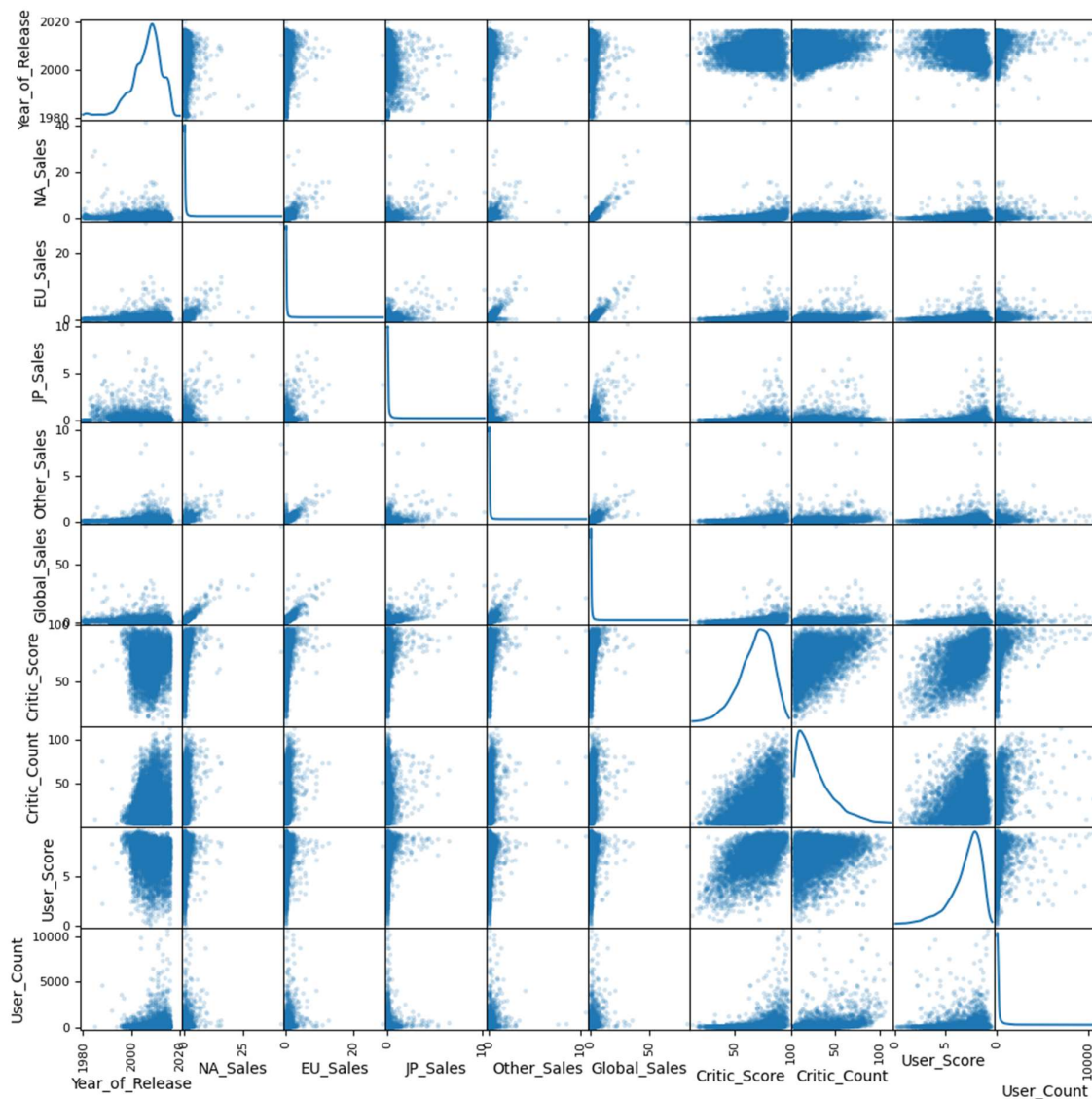
Slika 11 Horizontalni trakasti grafikon Publisher

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.



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.



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.

4. 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 avатарom, 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.

5. 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.

5.1. Google Gemini

1. Chat Bot

A chat application with an embeddable widget that can be used as a chatbot

App Input

Define the input fields you want this app to accept. These will be rendered as a form for users to fill out. If using the app via the API, the input fields will form the JSON schema for the input data.

Name	Description	Type	Options	Required	Actions
Pitanje	Ovdje unesite pitanje	string	Yes	<input checked="" type="checkbox"/>	<div>↑</div> <div>↓</div> <div>🗑️</div>

Add Field

Configuration

Slika 14 Prikaz strukture chatbot modela GEMINI

2. File Extractor

Extract text from file represented as data uri

Drag 'n' drop some files here, or click to select files. Maximum size of each file is 50 MB. Only 1 file(s) can be uploaded at once.

2k2_merged.pdf

×

The file to extract text from

File Data

The base64 encoded data of file

The base64 encoded data of file

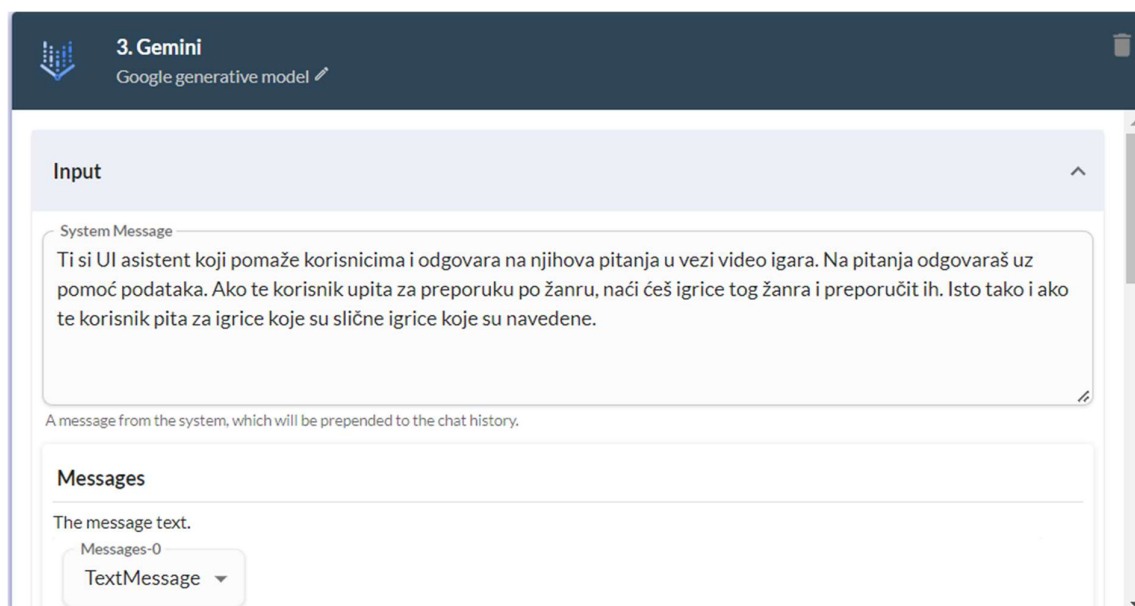
Query

[1] Input / pitanje

The query to search the document

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.



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.



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.

4. Application Output

Configure how the application will output data

Output Template

[3] Google / Gemini / content

Slika 18 Prikaz strukture chatbot modela GEMINI

5.2. 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.

3. ChatGPT

Takes a series of messages as input, and return a model-generated message as output

A list of messages, each with a role and message text.

Role

assistant

The role of the message sender. Can be 'user' or 'assistant' or 'system'.

Content

Context: [2] Promptly / File Extractor / text

Question: [1] Input / pitanje

The message text.

Name

The name of the author of this message or the function name.

Slika 19 Prikaz strukture chatbot modela OpenAI

4. Application Output

Configure how the application will output data

Output Template

[3] Open AI / ChatGPT / choices[0] / content

Slika 20 Prikaz strukture chatbot modela OpenAI

6. 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.

6.1. Razgovor s Google GEMINI

Chatbot za videoigre

Sviđa mi se igrice GTA V, možeš li preporučiti nešto slično?

Igre slične GTA V:

- **Red Dead Redemption 2:** Vestern akcija-avantura sa otvorenim svetom od Rockstar Gamesa, tvorca GTA V. Igra ima sličan otvoreni svet, detaljne likove i uzbudljivu priču.
- **Watch Dogs: Legion:** Akcija-avantura sa otvorenim svetom od Ubisofta koja se odvija u Londonu. Igra ima jedinstvenu mehaniku hakovanja i omogućava igračima da kontrolišu bilo kog NPC u gradu.
- **Sleeping Dogs:** Akcija-avantura sa otvorenim svetom od Square Enixa koja se odvija u Hong Kongu. Igra ima snažnu priču, borbe u blizini i vožnju vozilom.
- **Saints Row: The Third:** Akcija-avantura sa otvorenim svetom od Volitiona. Igra je poznata po svom humoru, preteranoj akciji i prilagodljivim likovima.
- **Mafia: Definitive Edition:** Remaster klasične akcije-avanture sa otvorenim svetom od Hangar 13. Igra ima poboljšanu grafiku, igru i priču.

Volim akcijske igre, možeš li mi preporučiti neke igrice tog žanra?



Preporučene akcijske igre:

- **Call of Duty: Modern Warfare II (2022):** Realistični FPS sa intenzivnim kampanjama i višegračkim režimima.
- **God of War Ragnarök (2022):** Akcijska avantura sa bogatim pričama i spektakularnim borbama.
- **Elden Ring (2022):** Open-world RPG sa izazovnim borbama i ogromnim svetom za istraživanje.
- **Apex Legends (2019):** Besplatna FPS igra sa herojima koji imaju jedinstvene sposobnosti i brzim tempom igre.
- **Fortnite (2017):** Besplatna igra preživljavanja sa izgradnjom i pucanjem u trećem licu.
- **Grand Theft Auto V (2013):** Kriminalistička akcija-avantura sa ogromnim otvorenim svetom i raznovrsnim misijama.
- **Doom Eternal (2020):** Brzotempovna FPS igra sa intenzivnim borbama i demonskim neprijateljima.
- **Assassin's Creed Valhalla (2020):** Akcijska RPG sa otvorenim svetom, vikinškim postavkama i stealth mehanikama.
- **The Witcher 3: Wild Hunt (2015):** Akcijska RPG sa otvorenim svetom, kompleksnim pričama i izazovnim borbama.
- **Metal Gear Solid V: The Phantom Pain (2015):** Akcijska avantura sa otvorenim svetom, stealth mehanikama i taktičkim borbama.



'ChatOutput' object has no attribute 'prediction'

Možeš mi navesti nešto o Namco studiju?



Namco Bandai Games

Osnovan: 1955. **Sjedište:** Tokio, Japan **Matična kompanija:** Bandai Namco Holdings Inc.

Poznate franšize:

- Pac-Man
- Tekken
- Soulcalibur
- Ace Combat
- Ridge Racer
- Tales of

Povijest:

Namco je osnovan 1955. godine kao tvrtka za iznajmljivanje mehaničkih konja. Tijekom 1970-ih, tvrtka je počela razvijati videoigre, a 1980. godine izdala je svoju prvu veliku uspješnicu, Pac-Man.

1990-ih, Namco se spojio s Bandai, što je rezultiralo osnivanjem Namco Bandai Gamesa. Tvrtka je nastavila razvijati i objavljivati popularne videoigre, uključujući Tekken i Soulcalibur.

1990-ih, Namco se spojio s Bandai, što je rezultiralo osnivanjem Namco Bandai Gamesa. Tvrtka je nastavila razvijati i objavljivati popularne videoigre, uključujući Tekken i Soulcalibur.

2005. godine, Namco Bandai se spojio s Namco Holdings, što je rezultiralo osnivanjem Bandai Namco Holdings Inc. Tvrtka je nastavila rasti i razvijati nove franšize, uključujući Ace Combat i Tales of.

Trenutno:

Namco Bandai Games je jedan od najvećih izdavača videoigara na svijetu. Tvrtka ima studije u Japanu, Sjedinjenim Državama i Europi. Namco Bandai Games je poznat po svojim visokokvalitetnim videoigrama i inovativnim pristupom dizajnu igara.



'ChatOutput' object has no attribute 'prediction'

Možeš mi navesti igrice s velikom ocjenom?



Akcijske igre

- Elden Ring (97/100 Metacritic)
- God of War Ragnarök (94/100 Metacritic)
- Horizon Forbidden West (88/100 Metacritic)

Avanturističke igre

- The Legend of Zelda: Breath of the Wild (97/100 Metacritic)
- Red Dead Redemption 2 (97/100 Metacritic)
- Ghost of Tsushima (83/100 Metacritic)

Role-playing igre

- Final Fantasy XIV: Endwalker (95/100 Metacritic)
- Elden Ring (97/100 Metacritic)
- The Witcher 3: Wild Hunt (93/100 Metacritic)

Simulacijske igre

- Microsoft Flight Simulator (90/100 Metacritic)
- Farming Simulator 22 (82/100 Metacritic)
- Cities: Skylines (85/100 Metacritic)

Simulacijske igre

- Microsoft Flight Simulator (90/100 Metacritic)
- Farming Simulator 22 (82/100 Metacritic)
- Cities: Skylines (85/100 Metacritic)

Strategijske igre

- Civilization VI (90/100 Metacritic)
- Total War: Warhammer III (88/100 Metacritic)
- Age of Empires IV (83/100 Metacritic)



'ChatOutput' object has no attribute 'prediction'

na osnovu igrice koju volim, koje platforme podržava?



Na osnovu igrice koju volim, koje platforme podržava?

Да бих одговорио на ово питање, потребна ми је информација о игрици коју волите. Молимо вас да ми је наведете.

6.2. Razgovor s OpenAI

Chatbot za video igre



Pozdrav

Sviđa mi se igrica GTA V, možeš li mi preporučiti nešto slično?

M



Ako ti se sviđa GTA V, mogu ti preporučiti nekoliko sličnih igara koje bi ti mogle biti zanimljive:

1. Red Dead Redemption 2 - Igra od istog developera kao GTA V (Rockstar Games) smještena u Divlji zapad. Ima sličan otvoreni svijet, duboku priču i mogućnost istraživanja.
2. Saints Row IV - Ova igra također spada u žanr open-world akcijskih igara s elementima vožnje i pucanja, ali s dodatnim nadrealnim elementima i humorom.
3. Sleeping Dogs - Igra koja se odvija u Hong Kongu i kombinira elemente vožnje, pucanja i borbe. Ima snažnu priču i otvoreni svijet za istraživanje.
4. Watch Dogs 2 - Ova igra kombinira akciju, vožnju i hakiranje u modernom urbanom okruženju, s fokusom na slobodnom istraživanju i interakciji s okolinom.

Nadam se da će ti ove preporuke pomoći pronaći igru koja ti odgovara! Ako te zanima više informacija o nekoj od ovih igara, slobodno pitaj.

Volim akcijske igre, možeš li mi preporučiti neke igrice tog žanra?

M



Ako voliš akcijske igre, evo nekoliko preporuka za igre u tom žanru koje bi ti mogle biti zanimljive:

1. Sekiro: Shadows Die Twice - Izrađena od strane FromSoftware, ova akcijska avantura kombinira intenzivnu borbu, izazovne protivnike i prekrasni svijet inspiriran srednjovjekovnim Japanom.
2. Devil May Cry 5 - U ovoj igri preuzimaš kontrolu nad demonolovcem Dantom koji se bori protiv nadnaravnih sila u brzom i stiliziranoj akcijskoj borbi.
3. Bayonetta - Akcijska igra u kojoj igraš kao čarobnica Bayonetta koja koristi svoje moći kako bi se borila protiv anđela i drugih neprijatelja. Ima brzu i fluidnu borbu s elementima hack-and-slash gameplaya.
4. Metal Gear Rising: Revengeance - Ova igra kombinira akcijsku borbu s elementima stealtha, a igrači preuzimaju ulogu Raidena, cyber-ninje koji se bori protiv moćnih neprijatelja.
5. Nier: Automata - RPG akcijska igra koja kombinira dinamičnu borbu s dubokom pričom o androidima koji se bore protiv mehaničkih horda na post-apokaliptičnoj Zemlji.

Nadam se da će ti ove preporuke pomoći pronaći akcijsku igru koja ti odgovara! Ako te zanima više informacija o nekoj od ovih igara, slobodno pitaj.

Možeš mi navesti nešto o Namco studiju?

M



Namco je japanski razvojni studio i izdavač video igara koji je osnovan 1955. godine. Kasnije je postao poznat kao Bandai Namco Entertainment nakon spajanja s tvrtkom Bandai. Namco je poznat po stvaranju klasičnih arkadnih igara poput Pac-Man, Tekken i Soulcalibur.

Osim arkadnih igara, Namco je također razvio i izdao razne druge igre u različitim žanrovima, uključujući RPG-ove, borilačke igre i arkadne klasike. Neki od njihovih popularnih serijala uključuju Tales serijal i Dark Souls (u suradnji s FromSoftware). Namco (sada Bandai Namco Entertainment) ima veliki utjecaj na industriju video igara i nastavlja stvarati kvalitetne naslove za igrače diljem svijeta.

Također, važno je napomenuti da je Namco jedan od pionira u industriji video igara, s dugom poviješću i naslijeđem koje je oblikovalo mnoge igrače i razvoj industrije.

Možeš mi navesti igrice s velikom ocjenom?

M



Evo nekoliko video igara koje su dobile visoke ocjene od kritičara i igrača:

1. The Legend of Zelda: Breath of the Wild - Ova igra za Nintendo Switch je dobila brojne pohvale zbog svoje otvorenosti, inovativnog dizajna svijeta i zanimljive mehanike igre. Smatra se jednom od najboljih igara svih vremena.
2. Red Dead Redemption 2 - Ova open-world igra od Rockstar Gamesa je dobila visoke ocjene zbog svoje duboke priče, realističnog svijeta i detaljne izvedbe. Igrači su pohvalili i karaktere te mehaniku igre.
3. The Witcher 3: Wild Hunt - RPG od CD Projekt Reda je osvojio brojne nagrade za svoju priču, likove, otvoreni svijet i kompleksne zadatke. Smatra se jednom od najboljih RPG igara u posljednjih nekoliko godina.
4. God of War (2018) - Nova verzija God of War igre za PS4 dobila je visoke ocjene zbog svoje priče, borbe, vizualnog dizajna i emotivne dubine likova. Igra je osvojila brojne nagrade.
5. Super Mario Odyssey - Ova igra za Nintendo Switch je dobila pohvale zbog kreativnosti, zabave i inovativnih mehanika igre. Smatra se jednom od najboljih platformskih igara.

Navedene igre su samo nekoliko primjera visoko ocijenjenih igara koje su uživale veliku popularnost među igračima i kritičarima. Ako te zanima više informacija o nekoj od ovih igara ili ako tražiš preporuke za specifičniji žanr, slobodno pitaj.





Iz tvog pitanja nije jasno koje je točno igrice koju voliš, pa ne mogu odrediti točne informacije o podržanim platformama. Međutim, općenito govoreći, većina modernih video igara podržava različite platforme, uključujući:

1. PC (Windows): Većina video igara je dostupna za PC platformu putem digitalnih trgovina poput Steam, Epic Games Store, Origin i sl.
2. PlayStation: Sonyjeva linija konzola uključujući PlayStation 4 (PS4) i PlayStation 5 (PS5) podržava širok spektar video igara.
3. Xbox: Microsoftova linija konzola uključujući Xbox One i Xbox Series X/S također podržava mnoge video igre, često dostupne putem digitalnih trgovina poput Xbox Store.
4. Nintendo: Nintendo Switch je popularna prijenosna konzola koja podržava mnoge ekskluzivne naslove Nintendo franšiza, ali i brojne neovisne i treće strane igre.
5. Mobilne platforme: Igre su također dostupne na mobilnim uređajima poput pametnih telefona i tableta putem aplikacija koje se mogu preuzeti iz App Storea (iOS) i Google Play Storea (Android).

Navedene su samo neke od platformi na kojima su video igre dostupne. Ako imaš konkretnu igru ili platformu na umu, slobodno navedi kako bih mogao pružiti preciznije informacije o podršci za tu igru.

7. 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.

8. 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 korištenja 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:

9. 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*. Techradar. <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*.

10. Popis slika

Slika 1 Stabilnost izvedbe LLM-a (preuzeto od autora Dao X)	8
Slika 2 Usporedba uspješnosti LLM-a i vijetnamskih studenata (Preuzeto od autora Dao X)	8
Slika 3 Prikaz tablice podataka	9
Slika 4 Prikaz tablice podataka	9
Slika 5 Prikaz atributa (iz google colab-a).....	10
Slika 6 Prikaz statistike brožani atributa (iz google colab-a)	11
Slika 7 Histogram Year_of_Release	12
Slika 8 Histogram NA_Sales.....	13
Slika 9 Histogram EU_Sales.....	14
Slika 10 Histogram JP_Sales	15
Slika 11 Horizontalni trakasti grafikon Publisher	16
Slika 12 Horizontalni trakasti grafikon Rating	17
Slika 13 Scatter plot matrix numeričkih vrijednosti.....	18
Slika 14 Prikaz strukture chatbot modela GEMINI	21
Slika 15 Prikaz strukture chatbot modela GEMINI	21
Slika 16 Prikaz strukture chatbot modela GEMINI	22
Slika 17 Prikaz strukture chatbot modela GEMINI	22
Slika 18 Prikaz strukture chatbot modela GEMINI	23
Slika 19 Prikaz strukture chatbot modela OpenAI	23
Slika 20 Prikaz strukture chatbot modela OpenAI	23