Proširenje velikih jezičkih modela sistemom za pronalaženje informacija

Milan Bojić [1]

***Sadržaj* — U prethodnom periodu došlo je do velikog razvoja sistema koja koriste velike jezičke modele. Ovi sistemi su pokazali velike mogućnosti u raznim scenarijima, ali su i demonstrirali i značajne nedostatke, kao što su halucinacije. U ovom radu će biti predstavljeno rešenje gde se koristi eksterna baza podataka koja bi trebalo da poboljša performanse jezičkih modela i reši postojeće probleme.**

***Ključne reči* — Veliki jezički modeli, Sistemi za pronalaženje informacija, Odgovaranje na pitanja otvorenog domena, Halucinacije**

I UVOD

Ovaj rad se fokusira na razvoj sistema koji razdvaja izvor znanja od generatora odgovora u velikim jezičkim modelima, i pretvara ih u posebne celine. Da bi razumeli zašto je ovo bitno, potrebno je razumeti kako veliki jezički modeli rade.

Obrada prirodnog jezika je grana računarstva, gde se obrađuje i analizira ljudski jezik. Ova grana je veoma važna, jer je ljudski jezik najčešći način komunikacije između ljudi, i ako želimo da računari budu u stanju da komuniciraju sa ljudima, onda je potrebno da računari budu u stanju da razumeju ljudski jezik. Tokom godina su se razvila različita rešenja za ovaj problem, i današnje rešenje je nastalo predstavljanjem Transformer arhitekture [1] i velikih jezičkih modela.

Transformeri su arhitektura neuralnih mreža koja koriste slojeve

[1] Računarski fakultet, Beograd, Srbija (email: mbojic@raf.rs)

pažnje (engl. attention) za pronalaženje zavisnosti između ulaznih i izlaznih podataka. Prvi modeli su korišćeni za mašinsko prevođenje, ali su se pokazali kao odlični za mnoge zadatke u obradi prirodnih jezika. Zavisno od zadatka, modeli prilagođavaju svoju arhitekturu, koje komponente sadrže i kako su povezane.

Iako su je veliki jezički modeli pokazali kao veoma uspešno rešenje, oni imaju i neke značajne nedostatke. Prvi nedostatak koji se može primetiti jeste kompleksnost modela, koji zahteva veliki broj parametara, i od prvih modela Transformera [1] do danas, broj parametara se povećao za nekoliko redova veličine. Ovo je problem, jer je potrebno dosta resursa da bi se trenirao model, a i da bi se model koristio u realnom okruženju. Ovo je veliki problem za nezavisne istraživače, pa se većina istraživanja radi na većim univerzitetima ili u velikim kompanijama. Manji timovi se najčešće oslanjaju na već postojeće modele i fino ih podešavaju za specifične zadatke.

Drugi nedostatak jeste upravljanje memorijom, gde za razliku od klasičnih baza podataka, gde se podaci čuvaju eksplicitno u memoriji, u velikim jezičkim modelima se podaci čuvaju implicitno u parametrima modela. Ovakvo čuvanje informacija može da dovede do problema sa radom jezičkog modela, jer se pri prolasku podataka kroz model ne zna koji će se parametri aktivirati za koje podatke. Ovo se može videti u nekoliko značajnih problema. Prvi problem jeste halucinacija [2], gde model vraća nepoželjne rezultate, kao što su besmisleni odgovori ili odgovori koji nisu u skladu sa kontekstom. Drugi problem jeste moderiranje sadržaja, gde da bi jezički model mogao da se koristi u realnom okruženju, potrebno je da se kontroliše sadržaj koji model vraća. Postoje nekoliko rešenja koja pokušavaju da reše ovaj problem [3], ali oni nisu idealni jer ne mogu eksplicitno da utiču na parametre modela, već samo na izlaz modela. I treći problem, sličan prethodnom, jeste dodavanje novih informacija u model, jer trenutno ne postoji način da se modelu doda nova informacija, a da se model ponovo ne trenira.

II VELIKI JEZIČKI MODELI DOPUNJENI SISTEMOM ZA PRONALAŽENJE INFORMACIJA

U prethodnom periodu se pokazalo da Veliki jezički modeli dopunjeni sistemom za pronalaženje informacija (engl. Retrival -augmented Large Language models, RALLM) mogu da poboljšaju performanse jezičkih modela, tako što za generisanje odgovora pored korišćenja ugrađenog znanja i ulaznog konteksta, koristi se i eksterna baza podataka [4]. Za zadati ulazni tekst model koristi sistem za

pronalaženje informacija za dohvatanje informacija relevantnih za ulazni tekst, i koristi generator za generisanje odgovora na osnovu ulaznog teksta i informacija.

To što je izvor informacija eksterni, to nam daje neke prednosti nad običnim jezičkim modelima, kao što su:

- Skalabilnost - smanjuje zahteve za veličinom modela i treniranjem, i omogućava lako proširenje baze podataka
- Preciznost - vezuje model za tačne informacije i smanjuje mogućnost Halucinacija
- Kontrola - daje nam mogućnost da kontrolišemo sadržaj koji model generiše
- Interpretabilnost - dovučene informacije služe kao referenca za generisani tekst

Veliki broj istraživanja je urađen u prethodnim godinama na ovu temu, sa naglaskom na primeni u jezičkim modelima, ali ima i istraživanja koja se bave sistemima sa više modela [5]. Sva istraživanja imaju istu osnovu, tačnije svi se mogu opisati kao *Veliki jezički modeli dopunjen k najbližih informacija iz sistema za pronalaženje informacija* (kNN-LLM) modeli, gde se glavne razlike svode na reprezentaciju informacija u vektorskom prostoru i na to koji LM model se koristi za generisanje izlaza.

Neki od primera istraživanja nad jezičkim modelima su: kNN-LM [6], DPR [7], RAG [8], REALM [9] i RETRO [10].

III Postavka eksperimenta

Za eksperimente je izabran zadatak odgovaranja na pitanja otvorenog domena (engl. Open-domain question answering). Eksperimentisano je sa nekoliko tipova modela, a za eksternu bazu podataka korišćen je **YottaAnswers sistem** [11], koji je napravljen za rešavanje ovog zadatka.

A. YottaAnswers

YottaAnswers je sistem za odgovaranje na činjenična pitanja na engleskom, koji je napravljen za rešavanje zadatka *Odgovaranje na pitanja otvorenog domena*. Sistem sadrži bazu od desetine milijardi odgovora na pitanja, koji su sakupljeni sa interneta, iz skupa podataka C4 [12]. Za upit sistemu se prosleđuje pitanje, a sistem vraća listu najboljih odgovora (format jednog odgovora je dat u Listingu 1).

Listing 1: Format odgovora YottaAnswers-a

```
{
      "answer": "Precizan odgovor na pitanje",
      "sentence": "Kontekst odakle je izvucen odgovor",
      "link": "Link ka izvoru odgovora"
}
```

Tačnije, korišćen je besplatan API čija je dokumentacija dostupna na linku[1].

B. Podaci

Za trening i test podatke su se koristili izlazni podaci iz YottaAnswers sistema, gde su se spajali pitanje i odgovori, tačnije svaki odgovor je bio kontekst sa naznačenim podskupom reči koje predstavljaju precizan odgovor. Na primer, na pitanje *Who wrote Crime and Punishment?* jedan od odgovora bi mogao biti:

Listing 2: Primer odgovora YottaAnswers-a na pitanje *Who wrote Crime and Punishment?*

```
{
      "answer": "Fyodor Dostoevsky",
      "sentence": "Fyodor Dostoevsky wrote Crime and Punishment",
      "link": "https :// en. wikipedia . org / wiki / Fyodor
      _Dostoevsky"
}.
```

U tom slučaju odgovor koji bi bio deo ulaza je **Fyodor Dostoevsky* wrote Crime and Punishment*. Izvučeno je 18 hiljada pitanja, gde su nekih 10% nasumična pitanja izvučena iz baze postavljenih pitanja, a ostatak su generisana na nekoliko tema (npr. Knjige-Pisci, Glavni gradovi država).

Podaci za treniranje su se onda generisali pomoću ChatGPT API-ja, gde je za svaki pasus (pitanje + 10 označenih odgovora) generisao jedan dugačak odgovor, koristeći samo ulazne podatke. Ova metoda gde se manji jezički model uči na primerima generisanim od strane većeg jezičkog modela se zove *destilacija* (engl. distillation) [13][14].

C. Modeli

Za eksperimente smo koristili dva modela: GPT-2 i Flan-T5. U ranijim istraživanjima se pokazalo da oba modela imaju dobre performanse na

[1]https://documenter.getpostman.com/view/21410829/UzBnpm9G

zadatku odgovaranja na pitanja otvorenog domena sa restrikcijama o kojima je pričano. Oba modela imaju zadatak da sumarizuju ulazni tekst tako da odgovor bude što koncizniji i da pre svega ima smisla.

GPT-2 je jezički model koji ima samo dekoder, koji je treniran na velikom broju tekstova sa interneta. Model je predstavljen od strane OpenAI-a 2019. godine [15]. Model ima više verzija sa različitim brojem parametara (do 1.5 milijardi parametara). Arhitektura koja uključuje samo dekoder je sa izlaskom sledećih generacija modela GPT-3 [16] i GPT-4 [17] postala veoma popularna, pa pošto je najbliži *open-source* model tim modelima GPT-2, njega koristimo za eksperimente. Arhitektura modela dozvoljava modelu da predviđa sledeću reč za zadati kontekst, bez nekog oblika razumevanja samog konteksta, što može da se pokaže kao problem za generisani izlaz.

Flan-T5 je koder-dekoder jezički model, koji je fino podešavana verzija originalnog T5 modela [18]. Model je predstavljen krajem 2022. godine od strane Google-a [19]. Model je fino podešavan na velikom broju zadataka, među kojima je i sumarizacija, i ima više verzija sa različitim brojem parametara (do 13 milijardi parametara).

Za eksperimente ćemo koristiti više modela sa različitim brojem parametara, da bismo videli kako se modeli ponašaju sa promenom veličine.

D. Evaluacija

Za evaluaciju koristimo metriku ROUGE (engl. Recall-Oriented Understudy for Gisting Evaluation) [20], skup metrika i softvera koji su specijalno namenjeni za evaluaciju automatski generisanih sažetaka u NLP-u. Metrika poredi automatski generisan sažetak sa referentnim sažetkom, i vraća vrednost između 0 i 1, gde je 1 najbolja vrednost.

Postoji nekoliko verzija ROUGE metrike:

- ROUGE-N - meri n-gram preklapanje između automatski generisanog sažetka i referentnog sažetka
- ROUGE-L - meri najdužu zajedničku podsekvencu između automatski generisanog sažetka i referentnog sažetka
- ROUGE-W - meri najdužu zajedničku podsekvencu gde se vrednuju uzastopne sekvence
- ROUGE-S - meri skip-bigram preklapanje između automatski generisanog sažetka i referentnog sažetka

Rouge metrika vraća tri vrednosti:

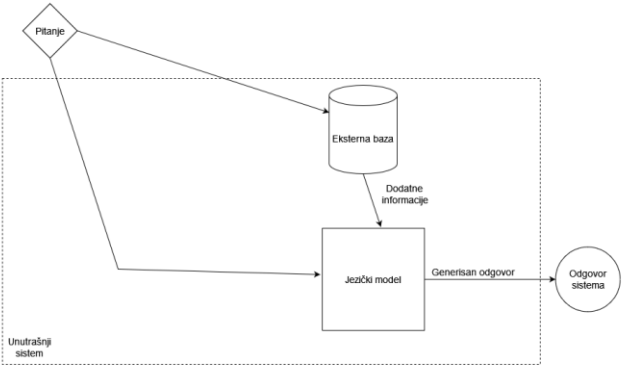
- Preciznost - procenat n-grama iz automatski generisanog sažetka koji se nalaze u referentnom sažetku
- Odziv - procenat n-grama iz referentnog sažetka koji se nalaze u automatski generisanom sažetku
- F-mera - harmonijska sredina preciznosti i odziva

E. Postavka za treniranje

Za treniranje je korišćen server sa 2x Nvidia Tesla V100 32GB GPU, 128GB RAM. Takođe da bi se ubrzalo treniranje, korišćena je Microsoft-ova biblioteka za optimizaciju **Deepspeed** [21], koja omogućava paralelizaciju treniranja na više GPU-ova.

F. Zadatak

Za razliku ranijih radova koji su se bavili ovom temom, model dobija podatke iz eksterne baze kao sam ulaz, i za izlaz se očekuje jedinstven odgovor na pitanje. Pošto su modelu dati svi potrebni podaci da bi se odgovorilo na pitanje, očekuje se da model bude u stanju da da tačan odgovor na pitanje. Zadatak modela može se posmatrati kao i pametna sumarizacija ulaznih podataka.



Sl. 1: Opšti dijagram sistema za odgovaranje na pitanja otvorenog domena

Na Sl. 1 je prikazan opšti dijagram sistema za odgovaranje na pitanja otvorenog domena.

Između generisanog odgovora i odgovora sistema, može postojati i dodatna obrada teksta, ispravljanje grešaka i slično, ali to nije predmet ovog rada.

IV Rezultati eksperimenta

Za eksperiment je korišćeno četiri različitih modela: GPT-2-medium (355 miliona parametara), GPT-2-large (774 miliona parametara), Flan-T5-Large (780 miliona parametara) i Flan-T5-XL (3 milijarde parametara). Takođe se radio eksperiment sa GPT-2-XL (1.5 milijardi parametara), ali je eksperiment prekinut zbog tehničkih problema.

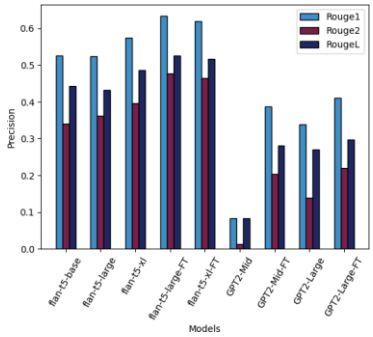
Za treniranje imamo 18.000 primera, gde se 17000 koristi za treniranje, a 1000 za validaciju. Za testiranje imamo 300 primera. Za treniranje svih modela je korišćen isti broj epoha (10). Za veličinu batch-eva (engl. batch size) se uz korišćenje akumulacije gradijenta koristila vrednost 32, odnosno broj primera nakon kojeg su se pravile korekcije parametara je bio 32. Tačni parametri su zavisili od modela i njegove veličine.

Rezultati eksperimenta su prikazani na Sl. 2. Korišćena je Rouge metrika, preciznije korišćeni su ROUGE-1, ROUGE-2 i ROUGE-L.

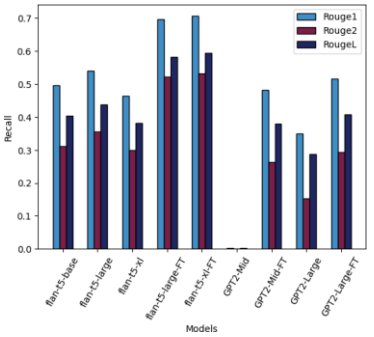
Ako prvo posmatramo rezultate za osnovne modele, vidimo da Flan-T5 modeli već imaju mogućnost da vrate prihvatljive rezultate (čak Flan-t5-base vraća prihvatljive rezultate). Sa druge strane, GPT-2 modeli nisu uspeli da vrate bilo kakve rezultate, GPT-2-medium obično vraća prazne odgovore ili odgovore koji se sastoje samo od nasumičnih karaktera (što utiče na prosečnu dužinu odgovora koja je prikazana na Sl. 3a); dok GPT-2-large uglavnom pon- avlja ulazne podatke. Ovo ponašanje GPT-2-large modela je dovelo do toga da rezultati koji se vide na Sl. 2 budu bolji nego što oni zapravo jesu. Svi ovi rezultati su bili i za očekivanje, jer je GPT-2 model treniran za generisanje teksta, dok je Flan-T5 model dotreniran da radi na zadacima (među kojima je i odgovaranje na pitanja i sumarizacija).

Ako posmatramo rezultate za dotrenirane GPT-2 modele, vidimo da su rezul- tati bolji nego za osnovne GPT-2 modele, ali i dalje nisu zadovoljavajući. Oba modela vraćaju odgovore, ali ako je odgovor manji od neke granice, onda model počinje da ga ponavlja. Ovo može da se vidi na Sl. 3b, gde se vidi da nezavisno od veličine ulaza u model, GPT-2 model vraća tekst slične dužine, dok ostali modeli vraćaju odgovore različite dužine (uglavnom prateći dužinu ulaza).

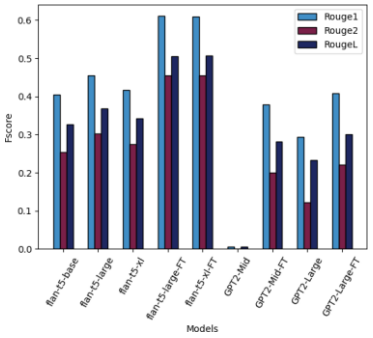
Sa druge strane, ako posmatramo dotrenirane Flan-T5 modele, vidimo da su rezultati bolji nego za osnovne Flan-T5 modele.



(a) *Rouge* metrika preciznosti među modelima



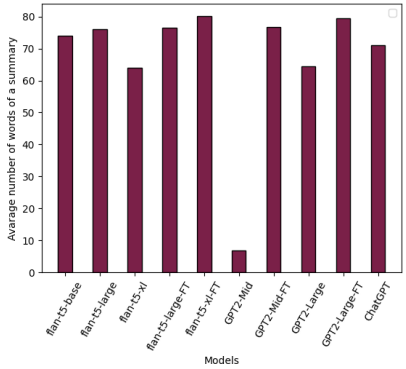
(b) *Rouge* metrika odziva među modelima



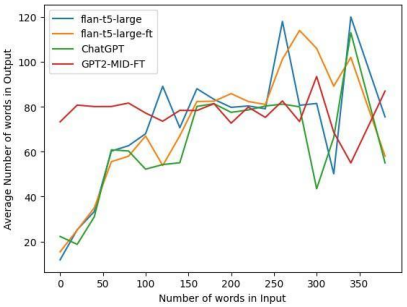
(c) *Rouge* metrika f-mere među modelima

Sl. 2: Rouge metrike među modelima

Napredak nije veliki, ali je dovoljan da se vidi da je model naučio da vraća odgovore koji su slični odgovorima iz trening skupa. Ovo se može objasniti time što je Flan-T5 model već dotreniran za zadatke koji su slični postavljenom zadatku, tako da je dotreniranje bilo samo da se model prilagodi obliku ulaznih podataka. Takođe, vidimo da Flan-T5-large i Flan-T5-XL modeli imaju gotovo iste rezultate, što najve- rovatnije znači da zadatak nije kompleksan i da je Flan-T5-large model dovoljan za ovaj zadatak. Rezultati bi se verovatno promenili da je zadatak bio kompleksniji, ili ako se modeli dotreniraju na većem broju zadataka.



(a) Prosečne dužine sažetka među modelima (sa ChatGPT izlazom)



(b) Prosečne dužine sažetka među modelima (sa ChatGPT izlazom) za zadati broj ulaznih reči

Sl. 3: Prosečna dužina sažetka među modelima (sa ChatGPT izlazom)

Par primera je ostavljeno u Dodatku A, sa pitanjem, izlazom YottaAnswers API-ja i izlazom svih modela u eksperimentu, kao i sa referentnom odgovorom.

V ZAKLJUČAK

Prvi zaključak koji se može izvući iz ovog eksperimenta je da je Flan-T5 model bolji za ovaj zadatak od GPT-2 modela, što je i očekivano, jer:

1. Flan-T5 model je izgrađen na arhitekturi koder-dekoder, koja je bolja za zadatke koji zahtevaju razumevanje konteksta, dok je GPT-2 model izgrađen na arhitekturi koja uključuje samo dekoder, i nema mogućnost implicitnog razumevanja konteksta.

2. Flan-T5 model je dotreniran na zadacima koji su slični posmatranom zadatku, dok je GPT-2 model treniran za generisanje teksta.

3. Flan-T5 model je mlađi model, to znači da su se za kreiranje modela mogle koristiti tehnike koje nisu bile dostupne pri kreiranju GPT-2 modela.

Drugi zaključak koji se može izvući je da veličina modela nije presudna za ovaj zadatak. Flan-T5-large i Flan-T5-XL modeli imaju gotovo iste rezultate, to znači da je Flan-T5-large model dovoljan za ovaj

zadatak. Isto važi i za GPT-2 model, gde GPT-2-medium i GPT-2-large modeli imaju gotovo iste rezultate. Ovo je i očekivano, jer je zadatak jednostavan, i ne zahteva veliki broj parametara.

Budući rad

Budući rad na ovu temu bi mogao da se fokusira na sledeće teme:

- Eksperimenti sa drugim modelima, iako je Flan-T5 mlad model (2022. godina), već postoje noviji modeli koji bi mogli da se isprobaju (npr. Llama, Llama2).
- Pronalaženje novih metrika za evaluaciju modela, koje bi bolje ocenjivale kvalitet odgovora.
- Mogle su se koristiti druge metode treniranja, kao što je *Reinforcement Learning from Human Feedback*.
- Modelovanje sistema za odgovaranje na različite vrste pitanja.
- Ubacivanje ovog pristupa u sisteme za dijalog (npr. ChatGPT).

Možda najveći napredak koji se može napraviti u sistemu jeste korišćenje drugog modela koji ima veću dužinu konteksta, Flan-T5 ima ograničenih 512 tokena, što kada se prevede u broj rečenica, znači da model može da razume samo 10-13 rečenica. Kada se gleda osnovni zadatak odgovaranja na pitanja otvorenog domena, ovo uglavnom neće praviti problem, ali ako se gleda neki kompleksniji zadatak koji očekuje da model razume veći kontekst, onda će biti problema. Trenuti predlog od dostupnih modela jeste Llama i Llama2 koji imaju veličinu konteksta od 2048 tokena, što je dosta veće od Flan-T5 modela.

LITERATURA

[1] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[2] Ziwei Ji et al. “Survey of hallucination in natural language generation”. In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.

[3] Todor Markov et al. “A holistic approach to undesired content detection in the real world”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12. 2023, pp. 15009–15018.

[4] Frank F Xu, Uri Alon, and Graham Neubig. “Why do Nearest Neighbor Language Models Work?” In: *arXiv preprint arXiv:2301.02828* (2023).

[5] Michihiro Yasunaga et al. “Retrieval-augmented multimodal language modeling”. In: (2023).

[6] Urvashi Khandelwal et al. “Generalization through memorization: Nearest neighbor language models”. In: *arXiv preprint arXiv:1911.00172* (2019).

[7] Vladimir Karpukhin et al. “Dense passage retrieval for open-domain question answering”. In: *arXiv preprint arXiv:2004.04906* (2020).

[8] Patrick Lewis et al. “Retrieval-augmented generation for knowledgeintensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[9] Kelvin Guu et al. “Retrieval augmented language model pre-training”. In: *International conference on machine learning*. PMLR. 2020, pp. 3929– 3938.

[10] Sebastian Borgeaud et al. “Improving language models by retrieving from trillions of tokens”. In: *International conference on machine learning*. PMLR. 2022, pp. 2206–2240.

[11] *Yottaanswers*. https://yottaanswers.com/.

[12] *C4 Dataset*. https://github.com/google-research/text-to- text-transfer-transformer. 2020.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[14] Rohan Taori et al. *Stanford Alpaca: An Instruction-following LLaMA model*. https : / / github . com / tatsu - lab / stanford _ alpaca. 2023.

[15] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[16] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[17] OpenAI. “GPT-4 Technical Report”. In: (2023).

[18] Colin Raffel et al. “Exploring the limits of transfer learning with a uni- fied text-to-text transformer”. In: *The Journal of Machine Learning Re- search* 21.1 (2020), pp. 5485–5551.

[19] Hyung Won Chung et al. “Scaling instruction-finetuned language models”. In: *arXiv preprint arXiv:2210.11416* (2022).

[20] Chin-Yew Lin. “Rouge: A package for automatic evaluation of summaries”. In: *Text summarization branches out*. 2004, pp. 74–81.

[21] microsoft. *Deepspeed*. https://github.com/microsoft/DeepSpeed.

A Dodatak A

Ova sekcija sadrži nekoliko primera i rezultata za svaki model u eksperimentu.

Tabela 1: Primer 1

Pitanje	When was Godels theory introduced?
Izlaz YottaAnswers API-ja	(These follow from the Completeness Theorem for predicate logic, Gödel **1931**.).
Flan-t5-base	1931
Flan-t5-large	Godel's theory was introduced in 1931.
Flan-t5-xl	Godels theory was introduced in 1931.
Flan-t5-large-ft	Godel's theory was introduced in 1931.
Flan-t5-xl-ft	Gödel's theory was introduced in 1931 with the Completeness Theorem for predicate logic.
GPT-2-mid	
GPT-2-mid-ft	Godel's theory was introduced in 1931 by Gödel. The Completeness Theorem for predicate logic was introduced in 1931 by Gödel. The theory of Godel's theory was introduced in 1931 by Gödel. The theory of Godel's theory was introduced in 1931 by Gödel. The Completeness Theorem for predi- cate logic was introduced in 1931 by Gödel. The theory of Godel's theory was introduced in 1931 by Gödel. The theory of Godel's theory was introduced
GPT-2-large	Theönyi (del31) introduced theorem that the (ple:forall x:xall) isa (xall:forall x:xall) for all x (x:forall x:xall) for all x (x :forall x:xall) for all x x :forall x:xall
GPT-2-large-ft	Godels theory was introduced in 1931. The Completeness Theorem for predicate logic was introduced in 1931. Theels theory is a theorem in predicate logic that states that there is no isomorphism between two elements there. is theorem in pred- icate logic that states that there is no isomorphism between two elements there. The theorem states that there is no iso- morphism between two elements there. The theorem states that there is no isomorphism between two elements there. The theorem states that there
ChatGPT	Gödel's theory was introduced in 1931.

TABELA 2: PRIMER 2

Pitanje	who are founders of apple
Izlaz YottaAnswers API-ja	Founded in April, 1976, Apple is now **one of the most famous Silicon Valley companies of all time and all time's founders Steve Jobs, Steve Wozniak and Ronald Wayne** are each icons and legends of their trade. Apple is a well recog- nised multinational Organisation **founded by Late Steve Jobs and Steve Wozniak** and is world wide known for of- fering better quality products with amazing features.amazing featuresis a pride owner of many new technological Gadgets. Which might as well be **Steve Jobs and Steve Wozniak, the founders of Apple**. **Apple's famous founders Steve Jobs and Steve Wozniak** needed to borrow $250,000 from banks to get both microloans's now-behemoth company off the ground. Noah Wyle and Joey Slotnick portray **Apple founders Steve Job and Steve Wozniak**. Apple is launched in the year 1976 by Apple's three founders, **Steve Jobs, Ronald Wayne and Steve Wozniak in the USA**.
Flan-t5-base	Steve Jobs and Steve Wozniak are each icons and legends of their trade. Apple is a well recognised multinational Organisa- tion **founded by Late Steve Jobs and Steve Wozniak** and is world wide known for offering better quality products with amazing features.amazing featuresis a pride owner of many new technological Gadgets. Which might as well be **Steve Jobs and Steve Wozniak, the founders of Apple**. Founded in April, 1976, Apple is now **one of the most famous Sili- con Valley companies of all time and all time's founders Steve Jobs, Steve Wozniak and Ronald Wayne** are each icons and legends of their trade. Apple is a well recognised multinational Organisation **founded by Late Steve Jobs and Steve Wozniak** and is world wide known for offering better qual- ity products with amazing features.amazing featuresis a pride owner of many new technological Gad

Flan-t5-large	**Steve Jobs, Steve Wozniak and Ronald Wayne** are each icons and legends of their trade. **Apple is a well recognised multinational Organisation **founded by Late Steve Jobs and Steve Wozniak** and is world wide known for offering better quality products with amazing features.amazing featuresis a pride owner of many new technological Gadgets.

Flan-t5-xl	Steve Jobs, Steve Wozniak and Ronald Wayne are the founders of Apple.
Flan-t5-large-ft	The founders of Apple are Steve Jobs, Steve Wozniak, and Ronald Wayne.
Flan-t5-xl-ft	The founders of Apple are Steve Jobs, Steve Wozniak, and Ronald Wayne. Apple was founded in April, 1976, by these three founders. Steve Jobs and Steve Wozniak needed to bor- row $250,000 from banks to get the company off the ground. Noah Wyle and Joey Slotnick portray Apple founders Steve Job and Steve Wozniak.
GPT-2-mid	" " " " " " "
GPT-2-mid-ft	Steve Jobs and Apple Jobs were only founders of Apple. many new technological gadgets were invented by Steve Jobs and Steve Wozniak. Steve Jobs and Steve Wozniak invented the first notable computer program. founders of Apple, including Steve Jobs, Wozniak, and Ronald Wayne, are known for their innovative.. Apple Jobs Ronald Wayne and Steve Wozniak founded Apple in April 1976.ozniak and Wozniak invented the first notable computer program. Steve Jobs and
GPT-2-large	** Jobs Jobs and Wozniak are are founders of Apple.** ** and andozniak is Steve Jobs Wozniak. **. TheSteve Jobs and Steve Wozniak are the founders of Apple.** ** The Paste article also of course points out that the company is owned by the company'sday of founding, the late Steve Jobs. The article also in to name the company's three founders, Steve Jobs, Steve Wozniak and Ronald
GPT-2-large-ft	Steve Jobs and Steve Wozniak are the founders of Apple. They founders of Apple, a with aknown history of innovation, Ap- ple, by to-date, has been Apple's most successful company. Founded in April, 1976, Apple is now one of the most famous Silicon Valley companies of all time. time founders of Ap- ple, Steve Jobs, Steve Wozniak, and Ronald Wayne are the founders of Apple. They founded Apple in April, 1976, in the USA.ni
ChatGPT	The founders of Apple are Steve Jobs, Steve Wozniak, and Ronald Wayne.

Abstract — In the previous period, there has been a great development of systems that use large language models. These systems have shown great potential in various scenarios, but have also demonstrated significant shortcomings, such as hallucinations. This paper will present a solution that uses an external database that is supposed to im- prove the performance of language models and solve existing problems.

Key words — Large language models, Information retrieval systems, Open domain question answering, Hallucinations

RETRIVAL AUGMENTED LARGE LANGUAGE MODELS

Milan Bojić