ChatGPT vs. DeepSeek: Prevođenje prirodnog jezika u SQL kod

Ivan Jovanović
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
ij20233020@student.fon.bg.ac.rs

Milica Škembarević
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
milica.skembarevic@fon.bg.ac.rs
0000-0003-0649-3005

Olga Jejić
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
olga.jejic@fon.bg.ac.rs
0000-0002-6594-6388

Marija Đukić
Fakultet organizacionih nauka, Univerzitet u Beogradu
Beograd, Republika Srbija
marija.djukic@fon.bg.ac.rs
0000-0002-1136-4278

***Apstrakt* - Veliki jezički modeli su privukli značajnu pažnju zbog sposobnosti generisanja smislenog i preciznog teksta. Prevođenje prirodnog jezika u SQL kod, kao i primena velikih jezičkih modela u ovoj oblasti, postaje sve popularnija tema među istraživačima. U radu se analiziraju performanse dva velika jezička modela, GPT-4o i DeepSeek R1, u generisanju SQL koda iz prirodnog jezika koristeći BIRD skup podataka. Fokus istraživanja je na proceni tačnosti generisanja SQL koda u zavisnosti od složenosti upita, s posebnim akcentom na složenost šema baza podataka. Rezultati pokazuju umerene performanse oba modela, pri čemu GPT-4o postiže tačnost od 60.17%, a DeepSeek R1 tačnost od 60.37%. Iako su ove vrednosti gotovo identične, detaljna analiza otkriva značajne razlike u performansama, pri čemu DeepSeek R1 pokazuje bolje rezultate u rešavanju izazovnijih upita, dok oba modela ostvaruju sličnu tačnost kod jednostavnijih upita. Istraživanje ukazuje na potrebu za detaljnijim testiranjima modela na upitima različite složenosti, kako bi se stekla preciznija slika o njihovim stvarnim sposobnostima.**

Ključne reči – NL-to-SQL, ChatGPT, DeepSeek, veliki jezički modeli, obrada prirodnog jezika

I. Uvod

Organizacije koriste relacione baze podataka za upravljanje i analizu velikih količina strukturiranih podataka, čineći ih ključnim delom poslovnih sistema [1]. Kako se obim podataka povećava, raste i potreba za efikasnim preuzimanjem i analizom podataka. Međutim, rad sa bazama podataka zahteva poznavanje odgovarajućeg jezika, kao što je struktuirani upitni jezik (SQL). Iako više od polovine profesionalnih programera (51%) koristi SQL jezik, svega četvrtina (23.6%) poseduje formalno obrazovanje u ovoj oblasti [2]. Zbog svoje tehničke prirode, SQL često predstavlja prepreku korisnicima bez odgovarajuće obuke, čime se stvara jaz između potrebe za podacima i znanja neophodnog za njihovo preuzimanje.

Prevođenje prirodnog jezika u SQL (NL-to-SQL) ima potencijal da u velikoj meri promeni pristup podacima omogućavajući korisnicima da pretražuju baze podataka koristeći prirodni jezik, bez potrebe za poznavanjem SQL-a ili šeme baze podataka [1]. Na taj način korisnici koji nisu tehnički obučeni mogu jednostavno pristupati i analizirati podatke, što može biti posebno korisno u oblastima poput poslovne inteligencije, korisničke podrške, ali i naučnog istraživanja.

Kako baze podataka postaju sve složenije, formulisanje odgovarajućih upita postaje sve izazovnije. Nedavno, veliki jezički modeli (LLM) su privukli značajnu pažnju zbog sposobnosti generisanja smislenog i preciznog teksta. Njihova primena može unaprediti obradu prirodnog jezika, jer se mogu koristiti za generisanje SQL koda na osnovu korisničkog pitanja i odgovarajuće šeme baze podataka. Zahvaljujući obuci na velikim skupovima podataka, LLM modeli bolje razumeju složene odnose između prirodnog jezika i šeme baze podataka u poređenju sa ranijim modelima [3], što im omogućava da generišu preciznije i sveobuhvatnije odgovore.

U radu su analizirane performanse velikih jezičkih modela za generisanju SQL koda na osnovu prirodnog jezika, uz analizu njihove sposobnost da odgovore na upite različitih nivoa složenosti. Za potrebe analize korišćen je skup podataka BIRD (BIg Bench for LaRge-scale Database Grounded Text-to-SQL Evaluation), koji omogućava verodostojniju procenu zahvaljujući složenijim šemama baza podataka koje vernije odražavaju kompleksnost stvarnih sistema.

Rad je organizovan na sledeći način: poglavlje 2 predstavlja pregled literature iz oblasti, dok su u poglavlju 3 predstavljeni korišćeni veliki jezički modeli. Metodologija istraživanja i opis testiranja, odnosno studije slučaja, prikazani su u poglavljima 4 i 5 respektivno. Diskusija rezultata, kao i identifikovana ograničenja, opisani su u poglavlju 6, a zaključak je dat u poglavlju 7.

II. Pregled literature

Problem prevođenja prirodnog jezika u SQL upite, kao i sve šira primena velikih jezičkih modela u toj oblasti, postaje predmet sve intenzivnijeg interesovanja istraživačke zajednice. Iako je broj publikacija na ovu temu još uvek relativno ograničen, istraživači ispituju različite pristupe korišćenju LLM-a kako bi unapredili proces prevođenja prirodnog jezika u SQL kod.

Prirodni jezik može biti nejasan jer korisnički upiti nisu dovoljno precizni. [4] su razvili sistem zasnovan na GPT-3 modelu kako bi poboljšali razumevanje upita na prirodnom jeziku i razrešili potencijalne nejasnoće. [5] su primenili

LLM modele kako bi pravilno "poravnali" SQL koda sa šemom baze podataka, a predloženi sistem su testirali na skupovima podataka Spider i BIRD. [6] su istraživali kako prilagođavanje šeme baze podataka može poboljšati performanse NL-to-SQL alata uvodeći opise vrednosti baze podataka koji prilagođeni LLM modelima. [7] navodi da performanse postojećih NL-to-SQL modela opadaju zbog nedostatka domensko-specifičnih podataka za domen za treniranje. Autori predlažu rešenje koje pojednostavljuje kreiranje skupova podataka za treniranje NL-to-SQL modela.

Autori [8] predlažu *Chain-of-Programs* pristup za NL-to-SQL, koji koristi Pandas za dekompoziciju složenih SQL upita na jednostavne operacije, čime omogućava open-source LLM modelima da postignu performanse slične GPT-4, uz niže troškove. U radu [9] su zbog domensko-specifičnih podataka razvijene posebne instrukcije za primenu GPT modela za NL-to-SQL zadatke. DBCopilot [10] je sistem namenjen generisanju SQL koda iz prirodnog jezika u okviru velikih i složenih baza podataka. Korišćenjem jednostavnog modela za upravljanje šemom i LLM modela za generaciju upita, omogućeno je skalabilno i precizno kreiranje SQL izraza u dinamičnim okruženjima.

U svetlu dosadašnjih nalaza, ovaj rad istražuje sposobnost velikih jezičkih modela da razumeju prirodni jezik u kontekstu upita različitih nivoa složenosti, sa ciljem generisanja tačnih SQL izraza za baze podataka sa kompleksnim šemama. Na ovaj način simulira se složenost stvarnih informacionih sistema, čime se obezbeđuje verodostojnija i praktično relevantna evaluacija performansi analiziranih modela.

III. VELIKI JEZIČKI MODELI

Zbog sposobnosti generisanja smislenog i preciznog teksta, veliki jezički modeli izazvali su značajan interes u istraživačkim krugovima i među korisnicima, posebno nakon pojave ChatGPT-a. Jedan od modela analiziranih u ovom radu je GPT-4o, koji je razvijen od strane OpenAI fondacije i postao dostupan javnosti u maju 2024. godine [11]. Funkcionalnosti vezane za napredno generisanje koda uvedene su ranije, sa verzijom GPT-4 u martu 2023. godine [12]. Drugi model koji je obuhvaćen analizom je DeepSeek R1, razvijen od strane kompanije DeepSeekAI, koji je postao dostupan u januaru 2025. godine [13]. Poboljšanja u generisanju složenog koda uvedena su sa verzijom DeepSeekCoder u januaru 2024. godine [14]. GPT-4o je odabran zbog svoje široke primene i velike rasprostranjenosti među korisnicima, dok je DeepSeek privukao pažnju istraživačke zajednice kao novo i perspektivno open-source rešenje.

IV. METODOLOGIJA

Performanse dva navedena modela analizirane su na skupu podataka BIRD (BIg Bench for LaRge-scale Database Grounded Text-to-SQL Evaluation). BIRD [15] predstavlja jedan od često korišćenih skupova podataka u istraživanjima na temu NL-to-SQL, a odlikuje ga prisustvo složenijih šeme baza podataka, čime bolje odražava realne izazove u transformaciji prirodnog jezika u SQL kod. Drugi popularni skupovi, kao što su Spider i WikiSQL, sastoje se uglavnom od jednostavnih baza podataka koje su kreirane namenski za istraživačke svrhe i ne oslikavaju u potpunosti složenost stvarnih baza podataka [16]. Dodatno, WikiSQL skup podataka je orijentisan na upite koji se odnose na pojedinačne tabele, bez mogućnosti testiranja sposobnosti modela u kontekstu relacija unutar složenijih baza podataka.

U okviru istraživanja analizirane su performanse dva velika jezička modela u generisanju SQL koda na osnovu prirodnog jezika. Evaluacija modela sprovedena je korišćenjem unapred pripremljenih pitanja iz BIRD skupa podataka, koji obuhvata upite različitih nivoa složenosti – od jednostavnih do složenih. Svaka instanca u skupu podataka sastoji se od pitanja formulisanog na prirodnom jeziku, konteksta određene baze podataka, kao i odgovarajućeg SQL upita, takozvani „zlatni“ kod. Za ocenu tačnosti generisanog koda korišćena je metrika Exact Match, a po uzoru na rad [17], primenjena je njena varijanta – Exact-Set-Match Accuracy. Ova metrika se uglavnom koristi u zadacima u kojima je očekivani izlaz skup vrednosti, a u radu je primenjena za određivanje stepena podudarnosti između generisanog SQL upita i „zlatnog“ koda. Metrika ima binarnu vrednost: rezultat 1 dodeljuje se kada generisani upit u potpunosti odgovara referentnom, dok rezultat 0 označava svako odstupanje. Zbog mogućnosti da za jedno pitanje postoji više ispravnih SQL upita, ova metrika omogućava precizniju procenu performansi modela.

A. Istraživačka pitanja

Cilj ovog rada je analiza performansi velikih jezičkih modela u generisanju SQL koda na osnovu prirodnog jezika (NL-to-SQL). Istraživanje obuhvata analizu dva modela: GPT-4o, koji je razvijen od strane kompanije OpenAI, i DeepSeek R1, koji je proizvod kompanije DeepSeekAI.

Rad daje odgovor na sledeća istraživačka pitanja:

1. Koji od dva analizirana jezička modela pokazuje bolje performanse u generisanju SQL koda na osnovu prirodnog jezika?
2. Kakve su performanse LLM modela za generisanje SQL upita različitih nivoa složenosti?

V. STUDIJA SLUČAJA

Performanse dva navedena modela ispitane su na skupu podataka BIRD. Po uzoru na [16], a radi smanjenja troškova i očuvanja reprezentativnosti, korišćen je uzorak koji obuhvata 10% podataka iz svake baze u razvojnom delu BIRD skupa podataka. Uzorak se sastoji od ukupno 113 upita, od kojih je 68 jednostavnih, 36 upita umerene složenosti i 9 izazovnih upita. Za interakciju sa LLM modelima korišćen je API pristup, pri čemu je Python skripta korišćena za automatizaciju procesa postavljanja upita. Svaki prompt upućen modelima sastojao se od instrukcija za model, korisničkog pitanja formulisanog na prirodnom jeziku i konteksta baze podataka na osnovu kojeg se generiše SQL kod. Kao odgovor, modeli su vraćali generisane SQL upite za svaki postavljeni zahtev.

VI. REZULTATI

Rezultati istraživanja pružaju koristan uvid u sposobnosti velikih jezičkih modela za generisanje SQL koda. Oba modela pokazuju zadovoljavajuće ukupne performanse, sa tačnošću od 60.17% za GPT-4o i 60.37% za DeepSeek R1. Iako su ove vrednosti na prvi pogled gotovo identične, detaljnija analiza pokazuje značajne razlike u načinu na koji svaki model odgovara na upite različitih nivoa složenosti.

Za jednostavne upite, oba modela pokazuju podjednaku uspešnost, uz blagu prednost na strani DeepSeek R1 modela (68.25% u odnosu na 67.64%). Ovo ukazuje na to da oba modela podjednako dobro rešavaju osnovne NL-to-SQL zadataka, poput direktnog izbora kolona i jednostavnih uslova filtriranja. Kod umerenih upita, koji obično uključuju složenije logičke operacije ili višestruke uslove u okviru WHERE klauzule, primetan je pad performansi kod oba modela. GPT-4o beleži nešto bolje rezultate (47.22%) u poređenju sa DeepSeek R1 modelom (44.44%). Pad performansi je donekle očekivan, obzirom da upiti zahtevaju dublje razumevanje semantike i konteksta, što može predstavljati izazov za velike jezičke modele. Kod izazovnih upita, DeepSeek R1 je značajno nadmašio GPT-4o, sa tačnošću od 71.42% naspram 55.55%. Ovakav rezultat može ukazivati da DeepSeek R1 uspešnije prepoznaje šablone i bolje se prilagođava jezičkim obrascima prisutnim u prirodnom jeziku. Nasuprot tome, lošije performanse GPT-4o modela mogu ukazivati na probleme generalizacije sa porastom složenosti zahteva.

Tabela 1. Rezultati testiranja

Složenost upita / LLM	Jednostavan	Umeren	Izazovan	Ukupne performanse
GPT-4o	67.64%	47.22%	55.55%	**60.17%**
DeepSeek R1	68.25%	44.44%	71.42%	**60.37%**

Neke od najčešćih grešaka koje su modeli pravili kod jednostavnih upita uključuju izostavljanje DISTINCT klauzule ili JOIN operacija. Kod umerenih upita, greške su se uglavnom odnosile na netačne uslove u WHERE klauzuli, spajanje pogrešnih tabela ili izbor neodgovarajućih kolona prilikom JOIN operacija. Iako su oba modela pokazala relativno visok nivo efikasnosti u rešavanju NL-to-SQL zadataka, njihova uspešnost varira u zavisnosti od složenosti upita. Samim tim, rezultati ističu značaj testiranja modela na upitima različite složenosti, kako bi se dobila jasniji uvid u stvarne mogućnosti I ograničenja.

A. Ograničenja

Važno je naglasiti potencijalna ograničenja u istraživanju koja mogu uticati na tumačenje rezultata. Dizajn prompta može uticati na rezultate i oblikovati odgovore modela. Takođe, obim uzorka koji je korišćen u evaluaciji može umanjiti mogućnost generalizacije zaključaka. Dodatno, sama prirode velikih jezičkih modela koji "rade" na principu „crne kutije“ ograničava mogućnost tumačenja njihovog ponašanja, naročito kod vlasničkih modela poput GPT-4o. Odsustvo uvida u mehanizme rezonovanja otežava identifikaciju razloga za uspeh ili neuspeh kod pojedinačnih upita, čime se onemogućava dublje sagledavanje rada modela.

VII. ZAKLJUČAK

U ovom istraživanju su upoređene performanse dva velika jezička modela GPT-4o i DeepSeek R1 u kontekstu njihove sposobnosti prevođenja prirodnog jezika u SQL kod. Za testiranje je korišćen BIRD skup podataka zbog složenosti šema baza podataka koje su dostupne u ovom skupu podataka. Iako su oba modela ostvarila slične ukupne rezultate, sa tačnostima od 60.17% za GPT-4o i 60.37% za DeepSeek R1, detaljna analiza je ukazala na značajne razlike u njihovoj sposobnosti da rešavaju upite različite složenosti. DeepSeek R1 je pokazao bolje performanse pri rešavanju izazovnih upita, dok su oba modela bila podjednako uspešna u rešavanju jednostavnijih zadataka. Dobijeni rezultati ukazuju da oba modela imaju zadovoljavajući potencijal za generisanje SQL koda, ali njihova efikasnost varira u zavisnosti od složenosti zadatka.

Iako su oba modela ostvarila zadovoljavajuće rezultate, istraživanje ukazuje na prostor za unapređenje daljim testiranjem na većem skupu podataka. Takođe dublje razumevanje mehanizama na kojima ovi modeli funkcionišu doprinelo bi interpretaciji rezultata, uzimajući u obzir njihovu nedostatak transparentnosti i ograničenu mogućnost direktnog uvida u operativne principe procesiranja. Dodatno, istraživanje generisanja SQL koda na srpskom jeziku predstavlja važnu temu za buduća istraživanja, što bi doprinosilo primeni ovih tehnologija u lokalnom kontekstu. Dugoročno, istraživački napori trebali bi biti usmereni ka unapređenju sposobnosti modela za rešavanje kompleksnih upita i poboljšanju razumevanja jezičkih i semantičkih specifičnosti u različitim domenima primene.

A. Budući rad

Buduća istraživanja uključiće evaluaciju performansi velikih jezičkih modela u generisanju SQL koda za baze podataka stvarnih poslovnih sistema, s ciljem procene njihove sposobnosti u razumevanju složenih i realističnih šema baza podataka. Takođe, jedan od mogućih pravaca daljeg istraživanja obuhvata analizu primene LLM modela u kontekstu NoSQL baza podataka, uzimajući u obzir ograničenu dostupnost odgovarajućih skupova podataka u ovoj oblasti. Dodatno, buduća istraživanja mogu uključiti ispitivanje performansi modela u generisanju upita na osnovu prirodnog jezika na srpskom jeziku, čime bi se unapredila lokalizacija i omogućila šira primena ovih tehnologija u domaćem okruženju.

ZAHVALNICA

Ovo istraživanje je finansirano od strane Fakulteta organizacionih nauka, Univerziteta u Beogradu.

LITERATURA

[1] A. B. Kanburoğlu and F. B. Tek, “Text-to-SQL: A methodical review of challenges and models,” Turkish J. Electr. Eng. Comput. Sci., vol. 32, no. 3, str. 403–419, 2024.

[2] Stack Overflow, “Stack Overflow Developer Survey 2024,” Stack Exchange Inc., https://survey.stackoverflow.co/2024, 2024.

[3] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, et al., “Recent advances in natural language processing via large pre-trained language models: A survey,” ACM Comput. Surv., vol. 56, no. 2, str. 1–40, 2023.

[4] A. Attawar, S. Vora, P. Narechania, V. Sawant, and H. Vora, “NLSQL: generating and executing SQL queries via natural language using large language models,” in Proc. Int. Conf. Adv. Comput. Technol. Appl. (ICACTA), pp. 1–6, okt. 2023.

[5] Y. Shen, X. Lin, J. Liu, Z. Huang, S. Wang, and Q. Liu, “SA-SQL: A Schema-Aligned Framework for Text-to-SQL through Large Language Models,” in Proc. Int. Conf. Comput. Linguist. Nat. Lang. Process. (CLNLP), pp. 71–77, jul. 2024.

[6] E. R. Nascimento, G. García, Y. T. Izquierdo, L. Feijó, G. M. Coelho, A. R. de Oliveira, et al., “LLM-Based Text-to-SQL for Real-World Databases,” SN Comput. Sci., vol. 6, no. 2, str. 130, 2025.

[7] Y. Tian, D. Lee, F. Wu, T. Mai, K. Qian, S. Sahai, et al., "Text-to-SQL Domain Adaptation via Human-LLM Collaborative Data Annotation," in Proc. 30th Int. Conf. Intell. User Interfaces, pp. 1398–1425, mart 2025.

[8] B. Xu, S. Li, Y. Wu, S. Wei, M. Du, H. Wang, and H. Song, "Chain-of-Program Prompting with Open-Source Large Language Models for Text-to-SQL," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), pp. 1–8, jun. 2024.

[9] G. Sun, R. Shen, L. Jin, Y. Wang, S. Xu, J. Chen, and W. Jiang, "Instruction tuning text-to-SQL with large language models in the power grid domain," in Proc. 4th Int. Conf. Control, Robot. Intell. Syst., pp. 59–63, avg. 2023.

[10] W. Zhang, Y. Shen, W. Lu, and Y. Zhuang, "Data-Copilot: Bridging Billions of Data and Humans with Autonomous Workflow," in ICLR Workshop on Large Language Model (LLM) Agents, 2024.

[11] OpenAI, "GPT-4o System Card," OpenAI, https://openai.com/index/gpt-4o-system-card/, 8. avg. 2024.

[12] OpenAI, "Chat Completions API," OpenAI, https://platform.openai.com/docs/guides/gpt/chat-completions-api, pristupljeno 11. apr. 2025.

[13] DeepSeek, "DeepSeek R1: Reasoning-focused language model," DeepSeek, https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf, pristupljeno 11. apr. 2025.

[14] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, et al., "DeepSeek-Coder: When the Large Language Model Meets Programming--The Rise of Code Intelligence," arXiv e-prints, arXiv:2401, 2024.

[15] J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, et al., "Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs," in Proc. 37th Int. Conf. Neural Inf. Process. Syst., pp. 42330–42357, dec. 2023.

[16] S. Talaei, M. Pourreza, Y. C. Chang, A. Mirhoseini, and A. Saberi, "CHESS: Contextual Harnessing for Efficient SQL Synthesis," arXiv e-prints, arXiv:2405, 2024.

[17] R. Zhong, T. Yu, and D. Klein, "Semantic evaluation for text-to-SQL with distilled test suites," in Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), pp. 396–411, nov. 2020.

ChatGPT vs. DeepSeek: Translating natural language into SQL code

Ivan Jovanović, Milica Škembarević, Olga Jejić, Marija Đukić

ABSTRACT

Large language models have attracted considerable attention for their ability to generate meaningful and precise text. The translation of natural language into SQL code, as well as the application of large language models in this area, is becoming an increasingly popular topic among researchers. This paper analyzes the performance of two large language models, GPT-4o and DeepSeek R1, in generating SQL code from natural language using the BIRD dataset. The focus of the research is on evaluating the accuracy of SQL query predictions based on query complexity, with a particular emphasis on the complexity of database schemas. The results show moderate performance for both models, with GPT-4o achieving an accuracy of 60.17% and DeepSeek R1 achieving an accuracy of 60.37%. Although these values are nearly identical, a detailed analysis reveals significant performance differences, with DeepSeek R1 demonstrating better results on more challenging queries, while both models perform similarly on simpler queries. The research highlights the need for further testing of the models on queries of varying complexity to obtain a more accurate picture of their true capabilities.