SVEUČILIŠTE U SPLITU

FAKULTET ELEKTROTEHNIKE, STROJARSTVA I BRODOGRADNJE

DOKTORSKI STUDIJ ELEKTROTEHNIKE I

INFORMACIJSKE TEHNOLOGIJE

KVALIFIKACIJSKI DOKTORSKI ISPIT

INTERAKCIJA SA SUSTAVIMA UMJETNE INTELIGENCIJE TEMELJENIM NA VELIKIM JEZIČNIM MODELIMA

Dino Nejašmić

Split, kolovoza 2024.

Sadržaj

[1 UVOD 3](#_Toc175671660)

[2 SUSTAVI UMJETNE INTELIGENCIJE 4](#_Toc175671661)

[2.1 Strojno učenje 5](#_Toc175671662)

[2.2 Generativna umjetna inteligencija 6](#_Toc175671663)

[3 VELIKI JEZIČNI MODELI 8](#_Toc175671664)

[3.1 Postavke velikih jezičnih modela 10](#_Toc175671665)

[3.2 Postavljanje upita za velike jezične modele 12](#_Toc175671666)

[3.2.1 Primjeri korištenja upita za velike jezične modele 14](#_Toc175671667)

[3.2.2 Tehnike postavljanja upita za velike jezične modele 18](#_Toc175671668)

[3.3 Usporedba velikih jezičnih modela 26](#_Toc175671669)

[3.3.1 Mjerila za usporedbu velikih jezičnih modela 26](#_Toc175671670)

[3.3.2 GPT-4 31](#_Toc175671671)

[3.3.3 Claude 3 Opus 36](#_Toc175671672)

[3.3.4 Gemini 1.5 Pro 38](#_Toc175671673)

[3.3.5 LLama 3 40](#_Toc175671674)

[3.4 Primjer korištenja upita za velike jezične modele: Interakcija s NAO robotom 42](#_Toc175671675)

[4 INTERAKCIJA SA SUSTAVIMA UMJETNE INTELIGENCIJE 44](#_Toc175671676)

[4.1 Taksonomija upotrebe umjetne inteligencije 46](#_Toc175671677)

[4.2 Umjetna inteligencija usmjerena na čovjeka 48](#_Toc175671678)

[4.3 Smjernice za oblikovanje interakcije sa sustavima umjetne inteligencije 51](#_Toc175671679)

[4.4 Agentski sustavi temeljeni na umjetnoj inteligenciji 53](#_Toc175671680)

[4.5 Bihevioristički pristup u analizi sustava umjetne inteligencije 56](#_Toc175671681)

[5 PREGLED PODRUČJA: INTERAKCIJA SA SUSTAVIMA UMJETNE INTELIGENCIJE TEMELJENIM NA VELIKIM JEZIČNIM MODELIMA 57](#_Toc175671682)

[5.1 Područje interakcije čovjeka i računala (engl. *Human Computer Interaction, HCI*) 62](#_Toc175671683)

[5.2 Područje oblikovanja upita za velike jezične modele (engl. *LLM Prompting*) 64](#_Toc175671684)

[5.3 Područje razvoja programske podrške (engl. *SW development*) 69](#_Toc175671685)

[5.4 Područje robotike (engl. *Robotics*) 69](#_Toc175671686)

[5.5 Područje obrazovanja (engl. *Education*) 70](#_Toc175671687)

[6 ZAKLJUČAK 72](#_Toc175671688)

[LITERATURA 74](#_Toc175671689)

[POPIS OZNAKA I KRATICA 84](#_Toc175671690)

[SAŽETAK 85](#_Toc175671691)

# UVOD

U posljednjih nekoliko godina mogućnosti sustava koji se temelje na umjetnoj inteligenciji (UI) eksponencijalno rastu i mijenjaju način na koji komuniciramo s tehnologijom i obavljamo svakodnevne zadatke. Posebno značajan napredak ostvaren je u području velikih jezičnih modela, koji su omogućili stvaranje sustava sposobnih za sofisticiranu i intuitivnu jezičnu interakciju s ljudima. Ovi modeli, temeljeni na dubokom učenju i velikim količinama tekstualnih podataka, otvaraju nove mogućnosti u brojnim domenama, od osobnih asistenata do naprednih alata za analizu i generiranje sadržaja. Međutim, s ovim napretkom dolaze i novi izazovi u oblikovanju interakcije između čovjeka i sustava UI. Ključni aspekt ove interakcije je oblikovanje upita za velike jezične modele, što predstavlja jedan od glavnih fokusa ovog istraživanja. Oblikovanje upita je vještina formuliranja instrukcija na način koji optimizira odgovore modela, omogućujući preciznije, relevantnije i korisnije rezultate. Ova tehnika nije samo pitanje sintakse, već zahtijeva duboko razumijevanje mogućnosti i ograničenja modela, kao i konteksta u kojem se koristi. Istražujući različite tehnike oblikovanja upita, ovaj rad nastoji pružiti uvid u najučinkovitije metode za iskorištavanje punog potencijala velikih jezičnih modela. Osim toga, u radu se istražuje kompleksnost interakcije sa sustavima UI temeljenim na velikim jezičnim modelima, analizirajući njihove mogućnosti, ograničenja i potencijalne primjene.

Ostatak rada strukturiran je na sljedeći način. Poglavlje 2 služi kao uvod u svijet umjetne inteligencije i njena uža područja koja su od interesa za ovaj rad počevši od strojnog učenja, preko neuronskih mreža i dubokog učenja, pa do generativne umjetne inteligencije, čime dolazimo do velikih jezičnih modela koji su zapravo u fokusu istraživanja.

Poglavlje 3, posvećeno je velikim jezičnim modelima. Najprije se daje uvid u osnovne postavke pomoću kojih se mogu podešavati veliki jezični modeli. Zatim se opisuju tehnike i metodologije za učinkovito postavljanje upita modelima te navode praktični primjeri korištenja takvih upita. Nakon toga slijedi analiza nekoliko najpoznatijih jezičnih modela i usporedba njihovih performansi uz opis mjerila koja se koriste za usporedbu modela. Poglavlje završava konkretnim primjerom korištenja velikih jezičnih modela kroz interakciju s NAO robotom.

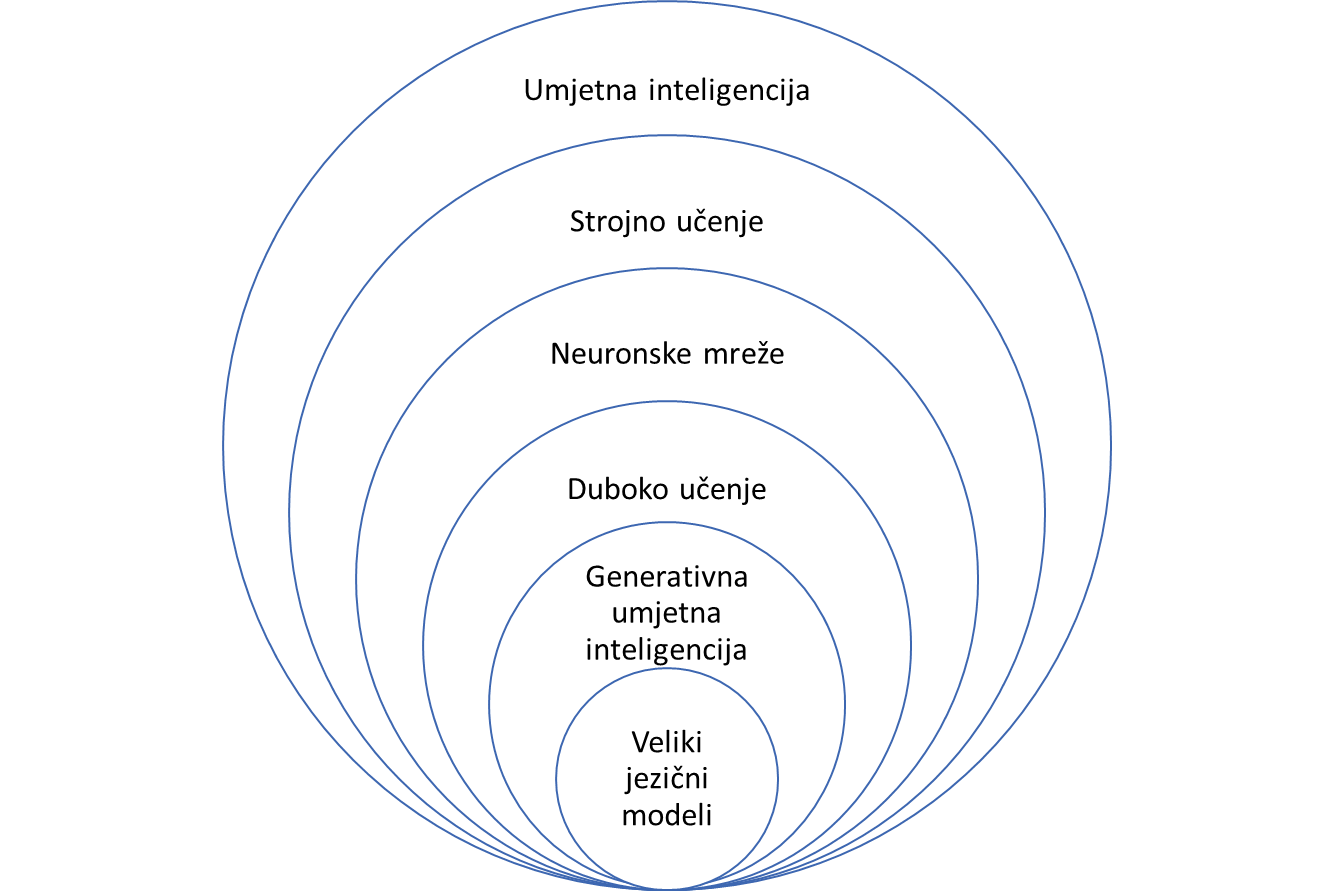
U poglavlju 4 opisana je interakcija sa sustavima UI, dana je taksonomija vezana uz korištenje UI, kao i smjernice za oblikovanje interakcije sa sustavima UI. Nadalje, napravljen je osvrt na umjetnu inteligenciju usmjerenu na čovjeka, agentske sustave temeljene na UI i bihevioristički pristup analizi sustava UI.

Konačno, poglavlje 5 donosi sustavan pregled relevantnih radova iz područja interakcije sa sustavima UI koji se temelje na velikim jezičnim modelima.

# SUSTAVI UMJETNE INTELIGENCIJE

Sustavi umjetne inteligencije (UI, engl. *Artificial Intelligence (AI) systems*) predstavljaju skup tehnologija, algoritama i računalnih programa koji omogućuju računalima simuliranje ljudske inteligencije i sposobnost rješavanja problema. Ovi sustavi se temelje na modelima koji mogu analizirati podatke, izvoditi zaključke, učiti iz iskustva te donositi odluke. Sustavi umjetne inteligencije su dizajnirani za rad s različitim razinama autonomije i mogu, za eksplicitne ili implicitne ciljeve, generirati rezultate kao što su predviđanja, preporuke ili odluke koje utječu na fizička ili virtualna okruženja. Imaju široku primjenu u industriji, medicini, financijama i brojnim drugim područjima [1].

Samostalno ili u kombinaciji s drugim tehnologijama poput senzora i robotike, sustavi temeljeni na umjetnoj inteligenciji mogu obavljati zadatke koji bi inače zahtijevali ljudsku inteligenciju. Primjeri takvih sustava u svakodnevnom životu uključuju digitalne asistente, autonomna vozila i alate generativne UI kao što je ChatGPT. Kao područje računalne znanosti, umjetna inteligencija među ostalim obuhvaća strojno učenje i duboko učenje, discipline koje se bave razvojem algoritama koji oponašaju procese donošenja odluka ljudskog mozga, omogućujući sustavima UI da "uče" iz dostupnih podataka i s vremenom postaju sve točniji u zadacima klasifikacije ili predviđanja. Današnji primjeri generativne umjetne inteligencije mogu učiti i sintetizirati ne samo ljudski jezik, već i druge tipove podataka poput slika, videa i programskog koda. Slika 2.1 prikazuje strukturu i odnose među pojmovima u području umjetne inteligencije koji su u fokusu ovog rada: strojno učenje, neuronske mreže, duboko učenje, generativna umjetna inteligencija i veliki jezični modeli [2].



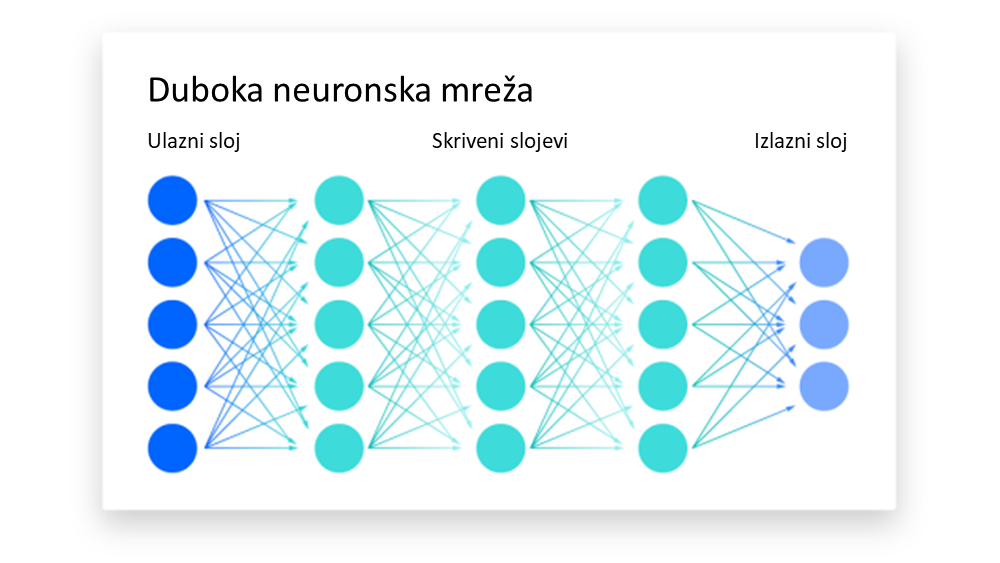
Slika 2.1 Struktura i odnosi u području umjetne inteligencije

Na slici je prikazana slojevita struktura u kojoj se svako sljedeće područje nalazi unutar šireg konteksta prethodnog područja. U daljnjem tekstu ovog poglavlja su objašnjeni osnovni koncepti, a obzirom da je fokus samog rada na velikim jezičnim modelima, njima je posvećeno zasebno poglavlje.

## Strojno učenje

Strojno učenje (engl. *Machine Learning, ML*) je područje umjetne inteligencije čiji je cilj postići da računala na temelju velike količine ulaznih podataka uočavaju zakonitosti i stvaraju modele koji se zatim primjenjuju za rješavanje zadataka. Proces učenja iz podataka zove se treniranje modela strojnog učenja. Taj proces uključuje optimizaciju parametara modela kroz iterativno prilagođavanje na temelju usporedbe vrijednosti predviđenih od strane modela s točnim vrijednostima iz skupa podataka za treniranje kako bi se smanjila pogreška predikcije. Strojno učenje je jedno od područja umjetne inteligencije koje se u posljednje vrijeme zasigurno najviše razvilo, pogotovo u pogledu rješavanja raznih kompleksnih zadataka koji se do sada nisu mogli riješiti na zadovoljavajući način. Tome je pridonio razvoj računala koja su sve brža, sposobna napraviti brojne numeričke operacije na kojima se strojno učenje i zasniva [3].

Jedni od složenijih modela strojnog učenja su umjetne neuronske mreže (engl. *Artificial Neural Network, ANN*) koje donose odluke na način sličan ljudskom mozgu, koristeći procese koji oponašaju rad bioloških neurona. Neuronske mreže oslanjaju se na podatke za treniranje kako bi učile i poboljšavale svoju točnost s vremenom. Svaka neuronska mreža sastoji se od slojeva čvorova ili neurona – ulaznog sloja, jednog ili više skrivenih slojeva, te izlaznog sloja. U slučaju kada neuronska mreža ima više od tri sloja, kao npr. mreža koju prikazuje Slika 2.2, smatra se dubokom neuronskom mrežom [4].



Slika 2.2 Duboka neuronska mreža, po uzoru na [4]

Svaki čvor je povezan s ostalima i ima vlastitu pridruženu težinu i prag. Ako je izlaz bilo kojeg pojedinog čvora iznad određenog praga, taj čvor se aktivira i šalje podatke u sljedeći sloj mreže. U suprotnom, podaci se ne prenose u sljedeći sloj mreže. Princip učenja umjetnih neuronskih mreža temelji se na prilagođavanju težina veza između neurona kroz proces povratnog širenja greške (engl. *backpropagation*) i optimizacije. Ovi procesi omogućuju neuronskim mrežama prepoznavanje određene pojave, usporedbu opcija i donošenje zaključka [4].

Takve višeslojne neuronske mreže, koje između ulaznog i izlaznog sloja imaju više skrivenih slojeva, temelj su dubokog učenja (engl. *Deep Learning*), metode koja se koristi u brojnim sustavima umjetne inteligencije. Postoje razna područja primjene metoda dubokog učenja kao što su: detekcija objekata na slikama, obrada prirodnog jezika, upravljanje autonomnim vozilima, otkrivanje financijskih prevara, dijagnosticiranje bolesti, itd.

Glavna prednost dubokog učenja je sposobnost sveobuhvatnog rješavanja zadataka. Ovaj pristup smanjuje potrebu za preliminarnom obradom podataka jer mreža samostalno identificira pravilnosti koje povezuju ulazni vektor s ciljnom varijablom. Mreža obavlja složen i naporan proces odabira značajki, što uvelike pojednostavljuje posao. Međutim, ove prednosti se ostvaruju samo uz dovoljno veliku količinu podataka za treniranje i pravilno odabranu arhitekturu neuronske mreže [5].

## Generativna umjetna inteligencija

Generativna umjetna inteligencija (engl. *Generative AI*) predstavlja područje umjetne inteligencije koje se bavi stvaranjem novog sadržaja na temelju različitih ulaznih podataka kao što su tekst, slike ili zvukovi. Generativni modeli UI koriste neuronske mreže kako bi identificirali obrasce i strukture unutar postojećih podataka te na temelju toga generirali nove i originalne sadržaje. Ovi modeli koriste različite pristupe učenju, uključujući polu-nadzirano i nenadzirano učenje, što omogućuje bržu i efikasniju obradu velikih količina podataka [6].

Generativni modeli omogućuju stvaranje novih podataka simulirajući raspodjelu podataka na temelju uzoraka naučenih iz skupa podataka. Postoje različite vrste generativnih modela koji pružaju različite prednosti ovisno o primjeni, zahtijevanoj kvaliteti i brzini generiranja rezultata, a neki od njih su [6]:

* Difuzijski modeli(engl. *Diffusion models*)– proces treniranja ovih modela uključuje dvije faze, najprije se dodaje slučajni šum podacima a zatim se taj šum uklanja kako bi se rekonstruirali podaci. Nakon uvježbavanja model može generirati nove podatke pokretanjem procesa uklanjanja šuma počevši od potpuno nasumičnog šuma. Difuzijski modeli omogućuju generiranje visokokvalitetnih rezultata ali je proces uvježbavanja često dugotrajan.
* Varijacijski autoenkoderi(engl. *Variational autoencoders, VAE*) – sastoje se od dvije neuronske mreže koje se nazivaju enkoder i dekoder. Enkoder pretvara ulazne podatke u komprimiranu reprezentaciju podataka koja sadrži informacije potrebne dekoderu za rekonstrukciju originalnih ulaznih podataka, a odbacuje nepotrebne informacije. Ova vrsta generativnih modela je brža u generiranju izlaza poput slika, ali njihova kvaliteta nije na razini koju pružaju difuzijski modeli.
* Generativne suparničke mreže (engl. *Generative adversarial networks, GAN*) – funkcioniraju tako da se dvije suparničke neuronske mreže međusobno natječu. Generator stvara nove primjere a diskriminator uči razlikovati stvarne od generiranih primjera. Postupak se ponavlja i na taj način se obje mreže kontinuirano unaprjeđuju svakom iteracijom. Ovaj pristup omogućuje brzo generiranje visokokvalitetnih izlaza, ali može imati ograničenu raznolikost generiranih primjera.
* Transformer – arhitektura koja se sastoji od dva glavna dijela: enkodera i dekodera. Enkoder prima ulaznu sekvencu i pretvara je u niz kontinuiranih reprezentacija. Dekoder zatim koristi te reprezentacije za generiranje izlazne sekvence. Ključna komponenta je mehanizam pažnje koji omogućuje modelu fokusiranje na različite dijelove ulazne sekvence pri generiranju svakog elementa izlazne sekvence. Transformer arhitekture postižu dobre rezultate u zadacima prevođenja jezika i obrade prirodnog jezika [7].

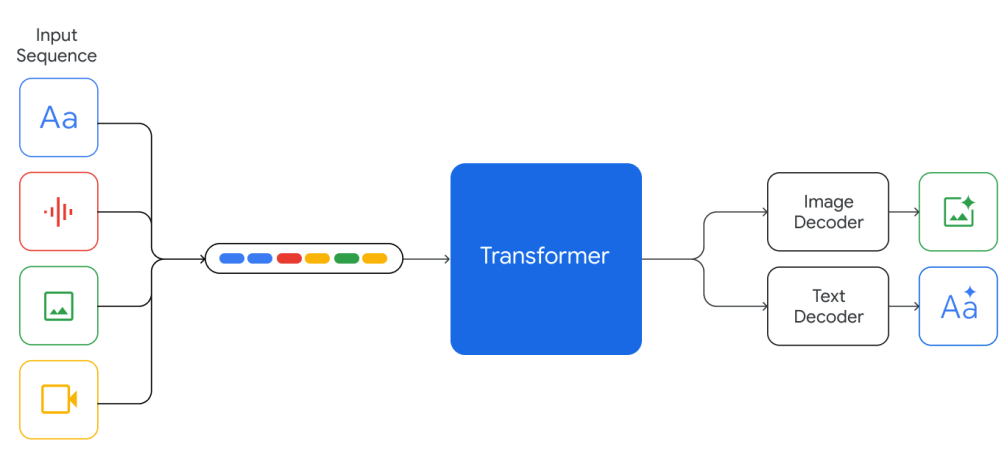
Primjene generativne umjetne inteligencije su razne, od stvaranja slika, audio i video zapisa, generiranja sintetičkih podataka za obučavanje modela umjetne inteligencije, do u posljednje vrijeme vrlo popularnih velikih jezičnih modela. S druge strane, postoje brojni izazovi poput složenosti i cijene potrebne računalne infrastrukture, brzine generiranja rezultata ili nedostatka kvalitetnih ulaznih podataka. Na početku sljedećeg poglavlja koje se bavi velikim jezičnim modelima opisani su još neke karakteristike generativne umjetne inteligencije odnosno transformer arhitekture.

# VELIKI JEZIČNI MODELI

Veliki jezični modeli (engl. *Large Language Model, LLM*) su vrsta modela umjetne inteligencije uvježbanih pomoću algoritama dubokog učenja za prepoznavanje, generiranje, prevođenje i sažimanje velikih količina pisanog ljudskog jezika i tekstualnih podataka. Ovi modeli spadaju jedne od najnaprednijih i najpristupačnijih rješenja za obradu prirodnog jezika (engl. *Natural Language Processing, NLP*) danas, sposobni ne samo procjenjivati postojeći tekst već i generirati originalan sadržaj na temelju korisničkih unosa i upita [8], [9].

LLM-ove karakterizira sposobnost razumijevanja i generiranja jezika sličnog ljudskom, često pomoću neuronskih mreža temeljenih na transformer arhitekturi koje paralelno obrađuju cijele sekvence osnovnih jedinica teksta, tzv. tokena, omogućujući upotrebu vrlo velikih modela sa stotinama milijardi parametara. To im omogućuje unos ogromnih količina podataka iz različitih izvora i učenje iz njih pomoću tehnika samostalnog učenja [8], [10].

Transformer arhitektura omogućuje modelima da obrade i generiraju prirodni jezik. Zasniva se na mehanizmima pažnje koji modelima daju sposobnost da razumiju kontekst i značenje riječi unutar dužih tekstova. Transformeri koriste pozicijsko kodiranje za očuvanje informacija o redoslijedu riječi, a obučavaju se na velikim skupovima podataka kako bi naučili predviđati sljedeću riječ u tekstu [7]. Slika 3.1 predstavlja jednostavni prikaz transformer arhitekture.



Slika 3.1 Jednostavan prikaz transformer arhitekture, preuzeto sa [11]

Ulazne sekvence (engl. *Input Sequence*) poput teksta, zvuka i slika se pretvaraju u vektorske reprezentacije (engl. *Embeddings*) koje predstavljaju svaku jedinicu ili token. Ovi vektori sadrže informacije o semantici i značenju tokena. Napredni algoritmi mogu interpretirati različite tipove podataka i dekodirati ih u druge oblike informacija ili medijskih izlaza, kao što prikazuje Slika 3.1. gdje ulazni podatci mogu biti tekst, zvuk, slika i video a izlazne podatke može dati dekoder slike (engl. *Image Decoder*) ili dekoder teksta (engl. *Text Decoder*)[11].

Tijekom obučavanja transformeri minimiziraju unakrsnu entropiju, mjeru koja kvantificira koliko dobro distribucija vjerojatnosti modela odgovara stvarnoj distribuciji riječi u korpusu. Optimizacija se obično provodi pomoću metode gradijentnog spusta, gdje se parametri modela postupno prilagođavaju kako bi se smanjila pogreška u predviđanju [12].

Jedna od ključnih prednosti transformer arhitekture je njena skalabilnost, što znači da performanse modela rastu s povećanjem broja parametara i veličine skupa podataka. Ovo svojstvo omogućuje obučavanje sve većih modela koji mogu obavljati sve složenije zadatke [8].

Proces uvježbavanja uključuje iterativno prilagođavanje vrijednosti parametara sve dok model ne predvidi ispravan sljedeći token iz prethodnog niza ulaznih tokena, a jednom kada je uvježban, model se može lako prilagoditi za obavljanje više zadataka koristeći relativno male skupove nadziranih podataka, proces poznat kao fino podešavanje (engl. *Fine Tuning*)[8].

Nadalje, LLM-ovi pokazuju sposobnost generalizacije i učenja u kontekstu, gdje model može izvoditi zadatke koji nisu izričito predstavljeni tijekom treninga. Primjerice, nakon što se modelu pruži nekoliko primjera ulazno-izlaznih parova, on može generirati odgovarajuće izlaze za nove ulaze, pokazujući razumijevanje i prilagodbu novim zadacima. Veliki jezični modeli generiraju odgovore na upite koje im zadaju korisnici, što mogu biti ljudi ili softver. Oblikovanje upita za LLM-ove zapravo je način na koji se "programira" model, obično davanjem smjernica ili nekih primjera kako uspješno izvršiti zadatak [13].

Neki od značajnih primjera LLM-ova uključuju GPT seriju modela razvijenu od strane tvrtke OpenAI, Google-ove Gemini modele, Meta-inu LLaMA obitelj i Claude modele tvrtke Anthropic, koji se koriste u raznim aplikacijama kao što su alati za razgovor (engl. *chatbot*) i alati za prevođenje jezika. Očekuje se da će se budućnost LLM-ova nastaviti razvijati, s potencijalnim napretkom u područjima kao što su multimodalni modeli koji mogu obrađivati tekst, ali i druge oblike informacija kao što su slike i zvuk [8], [9].

Korištenjem velikih jezičnih modela za generiranje teksta, mogu se izraditi aplikacije za:

* automatsko prevođenje tekstova između različitih jezika s visokom točnošću,
* generiranje ili unaprjeđivanje programskog koda na temelju prirodnog jezika,
* pomoć učenicima i studentima u obrazovanju npr. za rješavanje matematičkih problema ili objašnjavanje znanstvenih koncepata,
* interakcija s korisnicima pomoću naprednih chatbotova koji mogu voditi prirodne i korisne razgovore,
* razumijevanje objekata sa slike i davanje uputa za korištenje,
* odgovaranje na pitanja o bazi znanja,
* analizu rukom pisanih tekstova i izradu zapisnika,

te za razne druge primjene.

S druge strane, iako su veliki jezični modeli postigli značajan napredak u raznim područjima primjene, još uvijek postoje ograničenja poput nedostatka dugoročnog pamćenja, slabijeg razumijevanja konteksta i sklonosti stvaranju izmišljenih činjenica [8].

## Postavke velikih jezičnih modela

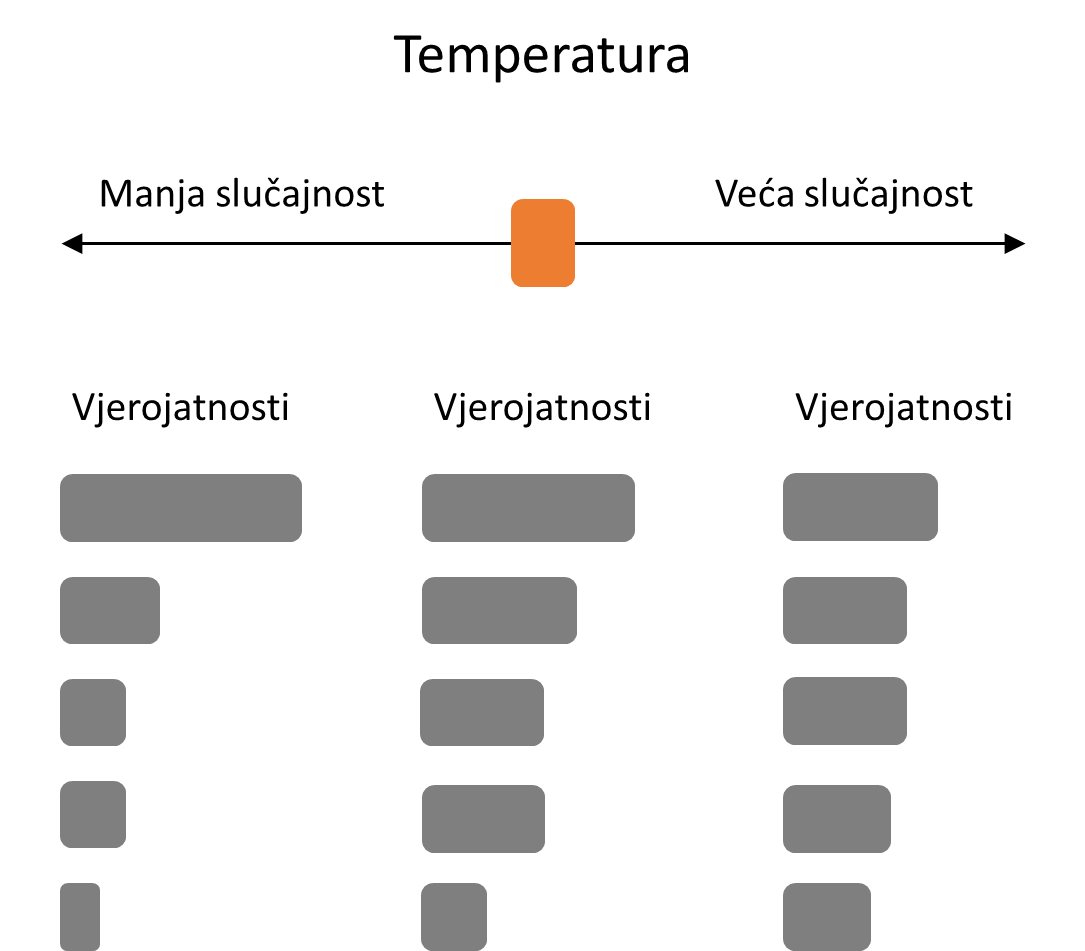
Prilikom dizajniranja i testiranja upita za jezične modela, ključno je pažljivo konfigurirati parametre kako bi se postigla željena razina pouzdanosti i raznovrsnosti u generiranim odgovorima. Ova prilagodba obično uključuje interakciju s modelom putem API-ja, gdje se manipulira određenim postavkama kako bi se utjecalo na način generiranja teksta [14].

Postavke velikih jezičnih modela koje su opisane niže u tekstu i koje se mogu podešavati prilikom interakcije s raznim modelima su:

* Temperatura
* Parametar Top P
* Duljina odgovora
* Sekvenca za zaustavljanje
* Kazna ponavljanja

Neki od modela omogućuju podešavanje sigurnosnih postavki na način da se prilagodi vjerojatnost dobivanja odgovora modela koji može sadržavati štetni sadržaj poput govora mržnje, opasnog sadržaja, seksualno eksplicitnog te uznemirujućeg sadržaja.

Jedan od ključnih parametara je temperatura (engl. *temperature*). Ova postavka određuje stupanj slučajnosti u odabiru sljedećeg tokena, kao što prikazuje Slika 3.2. Niža temperatura daje rezultate koji su više deterministički jer se uvijek odabire najvjerojatniji sljedeći token prema modelu. S druge strane, povećanje temperature potiče veću raznolikost u izlazima jer se povećava vjerojatnost odabira manje vjerojatnih tokena. Kako se povećava temperatura granica se smanjuje, dopuštajući modelu da koristi riječi sa manjom vjerojatnošću. Na primjer, za zadatke kao što su odgovori na pitanja temeljeni na činjenicama preporučuje se niža temperatura kako bi se potaknuli točni i koncizni odgovori, dok za generiranje pjesama ili kreativne tekstove može biti korisno povećati temperaturu [14], [15].



Slika 3.2 Utjecaj temperature na vjerojatnost odabira tokena, po uzoru na [15]

Drugi važan parametar je Top P, koji također omogućuje kontrolu determinističnosti modela, niža vrijednost Top P parametra rezultira točnijim odgovorima, dok se povećanjem ove vrijednosti potiče raznolikost u generiranim tekstovima. Na primjer, za precizne odgovore koriste se niže vrijednosti Top P, dok se za raznovrsnije odgovore koriste više vrijednosti. Top-p je sličan temperaturi, ali odabire najbolje tokene na temelju zbroja njihovih vjerojatnosti. Ne preporučuje se istovremeno koristiti Top P parametar i temperature parametar [14], [16].

Maksimalna duljina (engl. *Maximum Length*) također je važna postavka koja omogućuje kontrolu duljine generiranog odgovora. Ograničavanje maksimalne duljine pomaže u sprječavanju prekomjerno dugih ili irelevantnih odgovora te može utjecati na troškove obrade. Dodatno, korištenje sekvenci za zaustavljanje (engl. *Stop Sequences*) omogućuje kontrolu nad duljinom i strukturom generiranog odgovora. To mogu biti određene riječi, brojevi ili simboli kao što su razmak, novi red, točka ili zarez. Definiranjem određenih nizova znakova model se može zaustaviti pri generiranju, što je korisno za postizanje željenog formata odgovora [17].

Kazna frekvencije (engl. *Frequency Penalty*) i kazna prisutnosti (engl. *Presence Penalty*) su dodatne postavke koje utječu na ponavljanje riječi u generiranom tekstu. Kazna frekvencije smanjuje vjerojatnost ponovnog pojavljivanja već korištenih riječi i određuje se na temelju toga koliko se puta riječ ponovila u tekstu, dok kazna prisutnosti dodjeljuje jednaku kaznu za sve ponovljene jedinice, sprječavajući tako prekomjerno ponavljanje fraza [15].

Slika 3.3 prikazuje „*OpenAI playground*“ okruženje koje nudi različite opcije za interakciju s velikim modelima. Među ostalim opcijama za oblikovanje modela i interakcije, tu možemo vidjeti i prethodno opisane postavke velikih jezičnih modela.

A screenshot of a chat

Description automatically generated

Slika 3.3 OpenAI playground okruženje za oblikovanje interakcije s modelom

Pravilno podešavanje ovih parametara ključno je za postizanje željenih rezultata pri radu s jezičnim modelima. Razumijevanje i pravilna manipulacija ovim parametrima omogućuju optimizaciju performansi modela prema specifičnim zahtjevima i kontekstu primjene.

## Postavljanje upita za velike jezične modele

Postavljanje upita za velike jezične modele (engl. *LLM prompting*) je proces stvaranja i korištenja upita kako bi se usmjerio rad velikih jezičnih modela. Ova tehnika omogućuje korisnicima da dobiju željene rezultate od modela navođenjem specifičnih ulaznih podataka ili smjernica [14].

Upit (engl. *prompt*) se matematički može definirati kao funkcija koja oblikuje ulaz tako da se dobije željeni izlaz. U kontekstu velikih jezičnih modela, prompt je upit koji daje korisnik, a za koji je model oblikovan generirati odgovor. Funkciju preslikavanja možemo zapisati u obliku:

gdje je:

* 𝑃 funkcija prompta,
* 𝑥 predstavlja ulaz ili zahtjev koji se oblikuje funkcijom prompta,
* 𝑦 je izlaz ili odgovor koji generira model na temelju prompta.

Prompt se također može smatrati uvjetom koji je nužan da bi određeni ishod ili rezultat bio valjan, odnosno strukturirani upit koji daje smjernice modelu koji zadatak treba izvršiti ili koje informacije treba generirati. Oblikujući ulaz prompt direktno utječe na rezultat koji se dobije kao odgovor modela [18].

Dakle, prompt se može predstaviti kao funkcija koja pomoću skupa instrukcija, konteksta, ulaznih podataka i izlaznih indikatora inducira odgovor modela koji zadovoljava navedene zahtjeve. Ta se funkcija može zapisati na sljedeći način:

gdje je:

Instrukcija *(I)* - komponenta s instrukcijama specificira zadatak ili radnju koju bi model trebao izvršiti. Matematički, to se može predstaviti kao funkcija *I(x)* koja prima ulazni *x* i vraća skup instrukcija specifičan za zadatak.

Kontekst *(C) -* komponenta konteksta pruža vanjske informacije ili dodatne detalje koji usmjeravaju odgovor modela. To se može predstaviti kao funkcija *C(x)* koja prima ulazni *x* i vraća skup kontekstualnih informacija.

Ulazni podaci *(D)* - komponenta ulaznih podataka je određeni unos ili pitanje za koje se traži odgovor. Matematički, to se može prikazati kao funkcija *D(x)* koja prima ulazni *x* i vraća ulazne podatke.

Indikator izlaza *(O)* - komponenta indikatora izlaza specificira očekivani tip ili format izlaza. To se može predstaviti kao funkcija *O(x)* koja uzima ulazni *x* i vraća očekivani izlazni format.

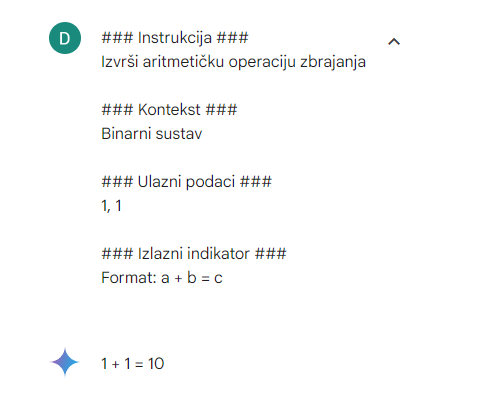
Kao primjer možemo uzeti rješavanje jednostavnih zadataka kao što je „Koliko je 1 + 1?“ kojeg prikazuje Slika 3.4. Odgovor na ovo pitanje je naizgled jednostavno dati, ali ako je pretpostavka da se tu radi o aritmetičkoj operaciji u dekadskom brojevnom sustavu zapravo predkoncepcija nastala kao rezultat učenja modela, možda nećemo dobiti željeni rezultat.

A white background with black text

Description automatically generated

Slika 3.4 Primjer jednostavnog upita za LLM

Međutim, ako upit postavimo tako da je komponenta instrukcije *I(x)* *= "Izvrši aritmetičku operaciju"*, komponenta konteksta *C(x)* *= "binarni sustav"*, komponenta ulaznih podataka *D(x)* *= "1, 1"*, a komponenta indikatora izlaza *O(x)* *= "Format: a + b = c"* funkcija prompta kombinirala bi ove komponente kao: *P(x) = I(x) ∘ C(x) ∘ D(x) ∘ O(x) = "Izvrši aritmetičku operaciju zbrajanja binarnih brojeva 1 i 1 te vrati rezultat u obliku a + b = c"*. Funkcija generiranja odgovora bi uzela ovaj upit i generirala odgovor: *f(P(x)) = „1+1= 10“*. Slika 3.5 prikazuje na ovaj način oblikovani upit te pruža okvir za razumijevanje načina na koji se može usmjeravati velike jezične modele da generiraju željene odgovore.



Slika 3.5 Oblikovanje upita za LLM

Pažljivom izradom komponenti instrukcija, konteksta, ulaznih podataka i indikatora izlaza, mogu se oblikovati učinkoviti upiti koji izazivaju točne odgovore velikih jezičnih modela [19].

Inženjerstvo postavljanja upita (engl. *prompt engineering*) relativno je nova disciplina za razvoj i optimizaciju upita u svrhu učinkovitog korištenja velikih jezičnih modela za širok raspon aplikacija i istraživačkih tema. Vještine koje se razvijaju u sklopu te discipline pomažu boljem razumijevanju mogućnosti i ograničenja velikih jezičnih modela. Međutim, to nije samo dizajniranje i razvoj prompta nego i širok raspon vještina i tehnika koje su korisne za interakciju s velikim jezičnim modelima, koje se mogu upotrijebiti za poboljšanje sigurnosti modela i izgradnju novih sposobnosti poput unaprjeđenja modela znanjem o specifičnoj domeni [14].

Postoje razne strategije postizanja boljih rezultata u interakciji s velikim jezičnim modelima a neke od njih su [19]:

* Pisanje jasnih uputa
  + Uključivanje detalja u upit kako bi se dobili relevantniji odgovori
  + Određivanje željene duljine i složenosti odgovora
  + Postavljanje modela u određenu ulogu
  + Ostavljanje što manje prostora za pogađanje
  + Jasno označavanje različitih dijelova upita
  + Određivanje koraka potrebnih za dovršetak zadatka
  + Navođenje primjera
* Usmjeravanje modela na pouzdane informacije
  + Usmjeravanje modela na odgovaranje citatima iz referentnog teksta
* Podjela složenih zadataka na jednostavnije podzadatke
  + Klasifikacija zadataka i određivanje relevantnih instrukcija
  + Sažimanje ili filtriranje prethodnih dugih dijaloga
  + Sažimanje dugih dokumenata po dijelovima i rekurzivna izrada punog sažetka
* Testiranje promjena na sustavan način
  + Definiranje sveobuhvatnog skupa testova

### Primjeri korištenja upita za velike jezične modele

U sljedećem dijelu navedeno je nekoliko primjera korištenja upita za velike jezične modele kako bi se pokazala mogućnost obavljanja različitih zadataka i predstavili neki od ključnih koncepata.

Sažimanje teksta

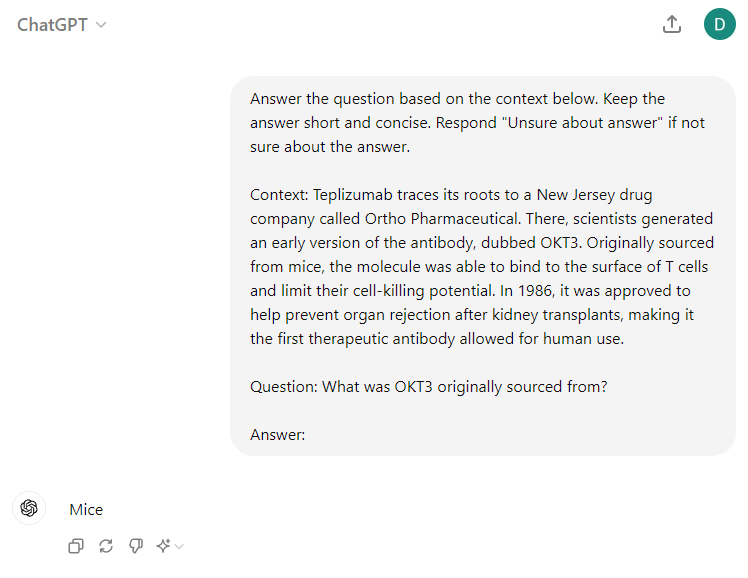
Jedan od osnovnih zadataka u obradi prirodnog jezika je sažimanje teksta odnosno proces prevođenja duljeg teksta u kraći oblik. Zadatak sažimanja teksta može uključivati ​​više različitih načina i domena. Jedna od važnijih primjena jezičnih modela je mogućnost sažimanja članaka i knjiga u sažetke koji se lako čitaju. LLM-ovi su se također pokazali učinkovitima u sažimanju poslovnih dokumenata kao što su npr. deklaracije proizvoda ili godišnja izvješća [17].

Izvlačenje Informacija

Izvlačenje informacija (engl. *Information Extraction*) je proces automatskog prepoznavanja i izdvajanja strukturiranih informacija iz neuređenih ili polustrukturiranih podataka. To može uključivati identificiranje određenih entiteta (npr. imena osoba, mjesta) i njihovih odnosa unutar teksta. U tom slučaju postavljeni upit kao ulazne podatke sadrži određeni tekst te instrukciju da se iz zadanog teksta izvuku točno određene informacije [20].

Odgovaranje na pitanja

Jedan od najboljih načina kako se može usmjeriti model da generira specifične odgovore je oblikovanje upita poput onoga što ga prikazuje Slika 3.6 Primjer oblikovanja upita za odgovaranje na pitanja [14]. U svrhu što boljeg obavljanja ovog zadatka upit se može podijeliti u sljedeće logičke komponente: instrukcije, kontekst, pitanje i frazu ("Odgovor:") kako bi se potaknulo model da započne generirati odgovor [21].



Slika 3.6 Primjer oblikovanja upita za odgovaranje na pitanja [14]

Klasifikacija teksta

Jedan od najčešćih oblika klasifikacije teksta je analiza sentimenta, koja dodjeljuje oznake poput "pozitivno", "negativno" ili "neutralno" nekoj tekstualnoj sekvenci. Npr. dajemo instrukciju modelu da klasificira recenziju filma [21].

Kao što je prethodno spomenuto, u svrhu dobivanja boljih rezultata upiti se mogu oblikovati pomoću instrukcija, konteksta, ulaznih podataka i indikatora izlaza. Ono što dodano poboljšava odgovore koje daju veliki jezični modeli su primjeri koji se daju kako bi se specificiralo ispravno ponašanje modela.

Razgovor

Jedna od najzanimljivijih primjena inženjerstva upita je davanje uputa sustavima velikih jezičnih modela (LLM) o njihovom ponašanju, namjeri i identitetu. Ovo je posebno korisno prilikom kreiranja sustava za razgovor, poput chatbotova za korisničku službu. Primjerice, možemo kreirati razgovorni sustav koji generira tehničke i znanstvene odgovore na pitanja. U tom procesu ključno je jasno definirati kako se sustav treba ponašati putem preciznih uputa. Ova tehnika, poznata kao poticanje uloga, omogućava chatbotu da preuzme određeni karakter ili stil komunikacije kako bi bolje odgovorio na specifične potrebe korisnika. S pravilnim usmjeravanjem, LLM sustavi mogu postati izuzetno učinkoviti alati za pružanje točnih i relevantnih informacija u različitim kontekstima [14].

Generiranje programskog koda

Pisanje ispravnog koda je vremenski zahtjevan posao u kojem se često događaju pogreške. Zbog toga se odavno razvijaju alati koji programerima pomažu npr. ukazivanjem na sintaksne pogreške, automatskim dovršavanjem koda u integriranim razvojnim okruženjima (IDE) ili različitim oblicima analize koda. Sljedeći korak u olakšavanju pisanja programskog koda je upotreba velikih jezičnih modela.

Generiranje koda je jedna od primjena u kojoj su LLM-ovi prilično učinkoviti. Copilot je izvrstan primjer za to. Postoji ogroman broj zadataka generiranja koda koji se mogu izvršiti pomoću dobro oblikovanih upita za LLM-ove kao npr. pisanje automatiziranih testova ili pisanje SQL upita za baze podataka.

Rasuđivanje

Jedan od izazova koji može biti izuzetno zahtjevan za velike jezične modele je onaj koji uključuje neki oblik zaključivanja. Rasuđivanje predstavlja zanimljivo područje zbog širokog spektra složenih primjena koje mogu proizaći iz takvih modela.

LLM-ovi ponekad teško rješavaju probleme koji su ljudima jednostavni, poput stvaranja akcijskih planova ili izvođenja složenog matematičkog ili logičkog rasuđivanja. Određeno poboljšanje u sposobnosti rasuđivanja postiže se zadavanjem većeg broja primjera ulaza i očekivanog izlaza, ali postojeći veliki jezični modeli i dalje nailaze na poteškoće pri obavljanju zadataka zaključivanja u više koraka, što zahtijeva primjenu još sofisticiranijih tehnika postavljanja pitanja. Neke od njih opisane su u sljedećem poglavlju [22].

Kao što je opisano u prethodnim primjerima, od modela se može tražiti da izvrši različite zadatke jednostavnim uputama koje ga usmjeravaju što da radi. To je važna osobina velikih jezičnih modela i mogućnost koju programeri sustava umjetne inteligencije već koriste za izradu vrlo efikasnih proizvoda. Veliki jezični modeli mogu se promatrati kao računala opće namjene koja izvršavaju programe određene upitima na prirodnom jeziku pozivima. Zbog toga postavljanje upita velikim jezičnim modelima možemo na neki način smatrati vrstom "programiranja" jer korisnici oblikuju upite kako bi model izvršio određene zadatke ili generirao određeni izlaz [23], [24].

Slika 3.7 prikazuje osnovni model funkcioniranja računalnih sustava u kojem ulaz (engl. *input*) predstavlja podatke ili informacije koje sustav prima, obrada (engl. *processing*) je manipulacija ili transformacija ulaznih podataka kako bi se dobili željeni rezultati, a izlaz (engl. *output*) su rezultati obrade koji se vraćaju korisniku ili se dalje koriste u sustavu [25].



Slika 3.7 Koncept ulaz-obrada-izlaz

Kao što kompajler može dati neočekivani izlaz kada se napravi pogreška u pisanju koda, isto tako i veliki jezični model može dati netočne ili neželjene rezultate ako mu se zada pogrešan ili nepotpun upit. U oba slučaja, greške u početnoj fazi mogu dovesti do nepredviđenih posljedica u završnom izlazu. Mogla bi se postaviti teza da je inženjerstvo postavljanja upita LLM-ovima prejednostavno da bi se smatralo programiranjem, međutim, slično kao što se u različitim programskim jezicima i paradigmama mogu razvijati raznoliki programi, od jednostavnog "Hello World" primjera do složenih aplikacija poput dubokih neuronskih mreža, tako i oblikovanje upita za velike jezične modele može biti vrlo složeno i sofisticirano. Inženjerstvo postavljanja upita velikim jezičnim modelima podrazumijeva pažljivo konstruiranje i prilagođavanje upita kako bi se postigli željeni rezultati. Upravo ta mogućnost finog podešavanja upita čine ovu disciplinu izazovnom.

Tehnike postavljanja upita velikim jezičnim modelima postaju sve popularnije jer omogućuju širok spektar aplikacija, uključujući generiranje teksta, stvaranje umjetnom inteligencijom potpomognutih alata za pomoć u radu, automatizaciju odgovora na upite i još puno toga. Neke od složenijih tehnika predstavljene su u sljedećem odjeljku.

### Tehnike postavljanja upita za velike jezične modele

Kao što je opisano u prethodnom odjeljku, poboljšanje postavljanja upita pomaže u postizanju boljih rezultata u različitim zadacima koje obavljaju veliki jezični modeli. U ovom odjeljku opisane su neke naprednije tehnike postavljanja upita koje nam omogućuju rješavanje složenijih zadataka te poboljšanje pouzdanosti i performansi modela, a to su:

* *Zero-Shot Prompting*
* *Few-Shot Prompting*
* *Chain-of-Thought (CoT) Prompting*
* *Self-consistency*
* *Prompt Chaining*
* *Tree of Thoughts (ToT)*
* *Retrieval Augmented Generation (RAG)*
* *Automatic Reasoning and Tool-use (ART)*
* *Automatic Prompt Engineer (APE)*
* *Active prompting*
* *Program-aided Language Models (PAL)*
* *ReAct (Reasoning and Acting)*
* *Reflexion*
* *Multimodal CoT*

Zero-Shot Prompting

Današnji veliki jezični modeli, poput GPT-3.5 Turbo, Gemini i Claude 3, podešeni su za praćenje uputa i obučeni na ogromnim količinama podataka. Zbog opsežne obuke, ovi modeli mogu izvršavati određene zadatke bez prethodnih primjera, što se naziva "*zero*-*shot* *prompting*". Zero-shot prompting podrazumijeva da upit korišten za interakciju s modelom ne sadrži primjere ili demonstracije, već izravno nalaže modelu da obavi zadatak bez dodatnih smjernica ili uputa [14].

Kako bi se dobili bolji rezultati pri korištenju Zero-Shot Prompting tehnike, odnosno kako bi modeli mogli bolje slijediti upute koje im se zadaju, koristi se ugađanje instrukcijama (engl. *instruction tuning*) za podešavanje i prilagodbu velikih jezičnih modela [26].

Osnovna ideja je sljedeća:

1. uzima se već prethodno obučeni veliki jezični model,
2. taj model se dodatno obučava (*fine tuning*) na skupu podataka koji se sastoji od uputa i željenih odgovora za te upute,
3. tijekom tog dodatnog obučavanja, model uči preslikavati zadane upute na ciljane izlaze, efektivno naučivši kako bolje slijediti različite vrste uputa.

Ključne prednosti ovog pristupa su:

1. omogućuje veću kontrolu nad ponašanjem modela sukladno zadanim uputama, umjesto nekontroliranog generiranja izlaza,
2. povećava sposobnost modela da obavlja različite zadatke definirane uputama, bez potpunog ponovnog treniranja od temelja,
3. učinkovitiji je od potpunog ponovnog treniranja jer se oslanja na znanje usvojeno iz početnog velikog modela.

Ugađanje instrukcijama se koristi za nadogradnju postojećih LLM-ova kako bi poboljšali svoj odgovor na upute, prilagodljivost raznim zadacima i slijeđenje ograničenja zadanih uputama. Mnogi od najsuvremenijih LLM asistenata danas koriste neku varijantu ovog pristupa. Međutim, pošto u nekim slučajevima zero-shot prompting ne daje dobre rezultate, preporučuje se putem postavljanja upita pružiti demonstracije ili primjere koji vode do nekoliko valjanih odgovora. U sljedećem odjeljku demonstrirano je tzv. navođenje s nekoliko primjera.

Few-Shot Prompting

Dakle, premda današnji veliki jezični modeli pokazuju izvanredne rezultate pri korištenju *zero-shot prompting* tehnike, oni još uvijek nisu uspješni u složenijim zadacima kada koriste *zero-shot* postavku. Instrukcije s nekoliko primjera mogu se koristiti kao tehnika za omogućavanje učenja u kontekstu gdje pružamo demonstracije u instrukcijama za usmjeravanje modela na bolju izvedbu. Demonstracije služe kao uvjet za sljedeće slučajeve u kojima želimo da model generira odgovor. Za teže zadatke možemo eksperimentirati s povećanjem broja demonstracija, npr. *3-shot, 5-shot, 10-shot* itd. [27].

Pri oblikovanju ovakvih upita važno je voditi računa o distribuciji primjera i formatu ulaznog teksta. Općenito gledajući, davanje većeg broja primjera može biti korisno za rješavanje nekih zadataka [28]. Međutim, kada *zero-shot* i *few-shot* tehnike ipak ne daju zadovoljavajuće rezultate, to može značiti da sve što je model naučio nije dovoljno za dobro obavljanje zadatka. Tada se preporučuje korištenje finog podešavanja modela ili eksperimentiranje s naprednijim tehnikama postavljanja upita. Jedna od tih tehnika je objašnjena u sljedećem odjeljku.

Chain-of-Thought (CoT) Prompting

U nekim situacijama model može davati pogrešne odgovore unatoč većem broju primjera koje je dobio kroz upit. U tom slučaju boljom se može pokazati tehnika postavljanja upita nazvana *Chain-of-Thought* koja potiče velike jezične modele da generiraju niz međukoraka ili razmišljanja koji vode do konačnog odgovora na složene zadatke. Ovaj pristup omogućuje modelima da razlože problem na manje dijelove, što olakšava razumijevanje i rješavanje problema. CoT se koristi za poboljšanje sposobnosti modela u rješavanju aritmetičkih, tzv. zdravorazumskih i simboličkih zadataka razmišljanja. Ova tehnika može se kombinirati s *few-shot prompting* tehnikom kako bi s dobili bolji rezultati na složenijim zadacima koji zahtijevaju obrazloženje prije odgovora [29].

Self-consistency

Samodosljednost (engl. *Self-consistency*) je pristup u kojem se modelu jednostavno postavlja isti upit više puta i uzima rezultat većine kao konačni odgovor. Smatra se nastavkom tehnike *Chain-of-Thought* i u kombinaciji s takvom vrstom upita daje bolje rezultate [30].

Prompt Chaining

Kako bi se poboljšala pouzdanost i performanse velikih jezičnih modela, jedna od ključnih tehnika postavljanja upita uključuje razbijanje zadataka na manje podzadatke. Nakon što su ti podzadaci definirani, LLM dobiva upit koji se odnosi na jedan podzadatak, a njegov se odgovor potom koristi kao ulaz za sljedeći upit. Ovaj pristup poznat je kao ulančavanje upita (engl. chaining), gdje se zadatak fragmentira na podzadatke s ciljem stvaranja niza međusobno povezanih upita. Ulančavanje upita korisno je za izvršavanje složenih zadataka koje bi model mogao teško riješiti čak i ako se od njega zatraži vrlo detaljan upit. Kod tehnike ulančavanja upita, lančani upiti izvode transformacije ili dodatne procese na generiranim odgovorima prije postizanja konačnog željenog stanja [14].

Osim postizanja boljih performansi, ulančavanje upita pomaže u povećanju transparentnosti aplikacija koje se oslanjaju na velike jezične modele, povećava mogućnost kontrole i pouzdanost. To znači da mogu puno lakše otkloniti pogreške u odgovorima modela te analizirati i poboljšati performanse u različitim fazama koje trebaju poboljšanja. Ulančavanje upita posebno je korisno kada se stvaraju pomoćnici za razgovor koje pokreće veliki jezični model te poboljšava personalizacija i korisničko iskustvo aplikacija [14], [31].

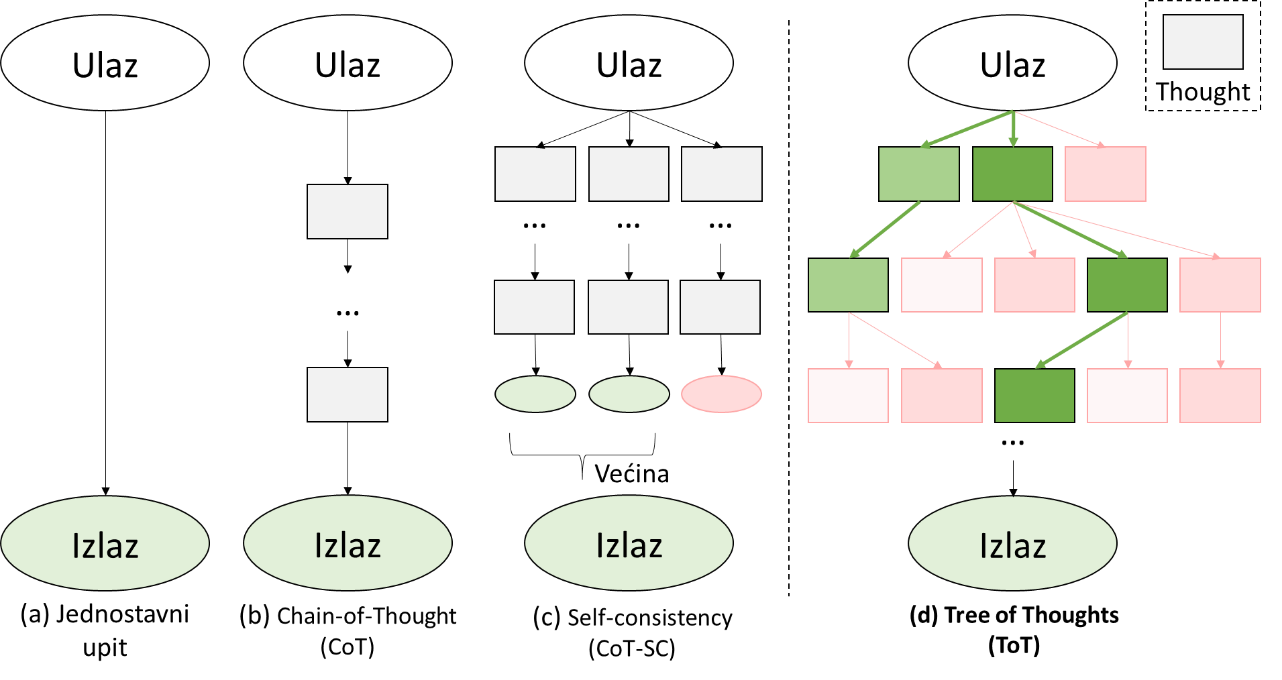
Ulančavanje upita može se koristiti u različitim scenarijima koji mogu uključivati ​​nekoliko operacija ili transformacija. Na primjer, jedan od uobičajenih slučajeva korištenja LLM-a uključuje odgovaranje na pitanja o velikom tekstualnom dokumentu. U tom slučaju mogu se dizajnirati dva različita upita gdje je prvi odgovor odgovoran za izdvajanje relevantnih citata za odgovor na pitanje, a drugi upit kao unos uzima citate i izvorni dokument za odgovor na dano pitanje. Pojednostavljenje problema i stvaranje lanaca upita koristan je pristup u kojem odgovori moraju proći kroz nekoliko operacija ili transformacija [31].

Tree of Thoughts (ToT)

Za složene zadatke koji zahtijevaju istraživanje ili strateško gledanje unaprijed, tradicionalne ili jednostavne tehnike poticanja nisu se pokazale uspješnima. Kao rješenje nudi se tehnika *Tree of Thoughts* koja proširuje CoT tehniku i omogućuje jezičnim modelima da istražuju različite putanje razmišljanja kako bi došli do najboljeg rješenja. Na ovaj način modeli „promišljeno“ donose odluke razmatrajući više različitih putanja razmišljanja i samo-ocjenjujući vlastite izbore odnosno odluke koje su donijeli. Ova tehnika značajno poboljšava sposobnost jezičnih modela u rješavanju problema koji zahtijevaju planiranje ili pretragu, kao što su npr. igre s brojevima, kreativno pisanje i mini križaljke [32].

Tehnika ToT inspirirana je načinom na koji ljudski um rješava složene zadatke razmišljanjem i pokušajima. Model pokušava tražiti rješenje na više načina što dovodi do grananja u više smjerova pretrage. Također je moguć povratak na prethodne korake u procesu razmišljanja i istraživanje drugih smjerova, što povećava “dugoročnu” sposobnost razmišljanja i rješavanja problema [33].

Slika 3.8 ilustrira tehniku ToT i njenu usporedbu s nekim od ostalih tehnika postavljanja upita koje su prethodno opisane u ovom poglavlju.



Slika 3.8 Različiti pristupi rješavanju problema pomoću LLM-ova, po uzoru na [32]

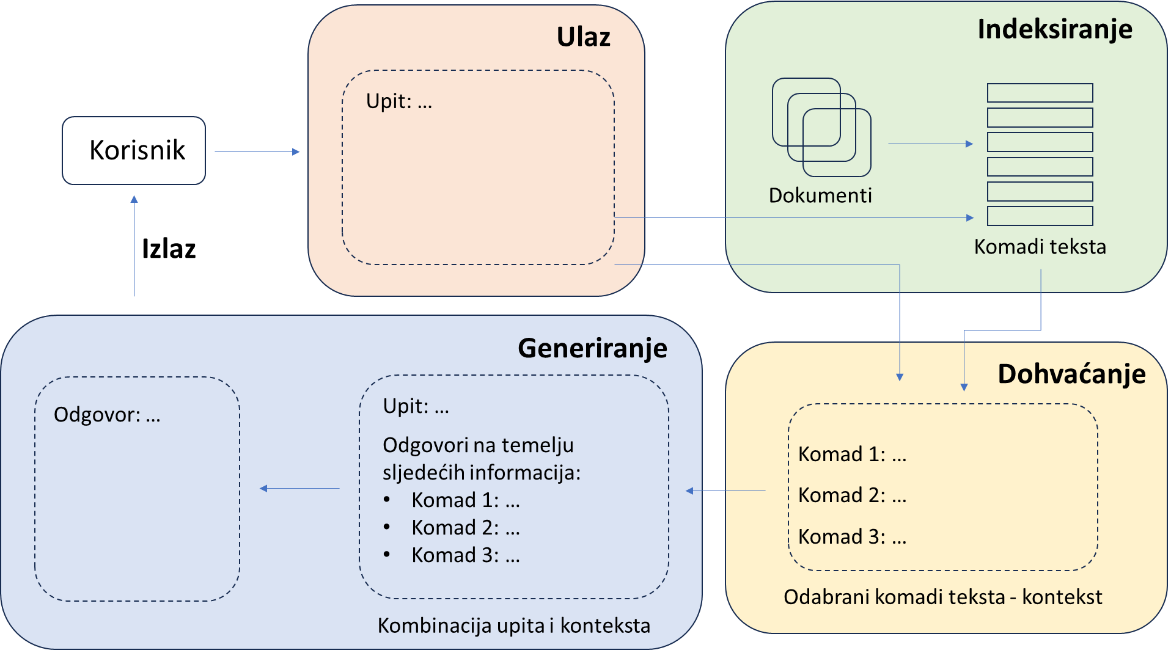
Kada koristi tehniku ToT model izgrađuje stablo misli, gdje misli predstavljaju koherentne jezične sekvence koje služe kao međukoraci prema rješavanju problema. Ovaj pristup omogućuje LLM-u da sam procijeni napredak kroz određene grane i zaključi vode li određene misli prema rješavanju problema. Ova sposobnost LLM-a da generira i evaluira svoje misli zatim se kombinira s algoritmima pretraživanja (npr. pretraživanje po širini) kako bi se omogućilo sustavno istraživanje misli s pogledom unaprijed i praćenjem unazad [14].

Retrieval Augmented Generation (RAG)

LLM-ovi napravljeni za opću upotrebu mogu se prilagoditi za postizanje boljih rezultata prilikom obavljanja specijalnih zadataka. Za složenije zadatke, moguće je izgraditi sustav temeljen na jezičnom modelu koji pristupa vanjskim izvorima znanja kako bi obavio te zadatke. To omogućuje povećanje dosljednosti i pouzdanosti generiranih odgovora te ublažavanje problema "halucinacije". Tehnika RAG kombinira komponentu za dohvaćanje informacija s modelom za generiranje teksta. Može se dodatno prilagoditi koristeći *Fine tuning*, a interno znanje modela može se nadograditi bez potrebe za ponovnim učenjem cijelog modela [14].

Ova tehnika kombinira znanje stečeno tijekom obuke modela s podatcima iz dohvaćenih dokumenata. Prednost ovakvog pristupa je mogućnost ažuriranja znanja modela bez potrebe za ponovnim treniranjem. To model čini prilagodljivim za situacije kada bi se činjenice mogle mijenjati kroz vrijeme. Ovo je vrlo korisno jer je parametarsko znanje velikih jezičnih modela statično. RAG omogućuje jezičnim modelima da zaobiđu ponovno učenje, omogućujući pristup najnovijim informacijama za generiranje pouzdanih izlaza putem generiranja temeljenog na dohvaćanju [34].

Slika 3.9 pokazuje sustav temeljen na RAG tehnici postavljanja upita.



Slika 3.9 Sustav temeljen na RAG tehnici postavljanja upita, po uzoru na [35]

Sustav prima upit i dohvaća skup relevantnih dokumenata iz izvora. Dokumenti se spajaju i zajedno s originalnom ulaznom naredbom tvore kontekst. Daljnji postupak se sastoji od tri glavna koraka. Najprije se podaci iz dokumenata spremaju u obliku malih komada teksta. Zatim se odabiru najrelevantniji komadi u skladu sa postavljenim upitom. Odabrani komadi teksta se zajedno s originalnom ulaznom naredbom prenose generatoru koji daje konačni izlaz [35].

Automatic Reasoning and Tool-use (ART)

ART je tehnika postavljanja upita koja omogućuje velikim jezičnim modelima da automatski generiraju tzv. korake razmišljanja kao program, koristeći postojeće primjere. Koristi se za rješavanje složenijih zadataka pomoću primjera višestrukih koraka razmišljanja i upotrebe vanjskih alata poput tražilica i alata za izvršavanja koda.

Tijekom izvršavanja može se zaustaviti generiranje rezultata kad god se pozivaju vanjski alati, nakon čega se integrira njihov izlaz, a zatim nastavlja s generiranjem. To omogućuje modelu da generalizira od danih primjera do dekompozicije novog zadatka i upotrebe alata na odgovarajućim mjestima. Zbog toga ART tehnika postiže značajno bolje rezultate na standardnim testovima u odnosu na *few-shot* i CoT tehniku, pogotovo na zadacima koji zahtijevaju aritmetičko ili algoritamsko rezoniranje. Ovaj okvir je proširiv i omogućuje korisnicima da poboljšaju performanse ispravljajući greške u specifičnim programima ili jednostavno dodavanjem novih alata u postojeće biblioteke [36].

Automatic Prompt Engineer (APE)

APE je algoritam koji automatski generira i odabire upite za velike jezične modele, s ciljem poboljšanja njihove izvedbe na različitim zadacima. Algoritam tretira upite kao “program” koji se optimizira pretraživanjem skupine kandidata za upit koje predlaže LLM, kako bi se maksimizirala odabrana funkcija ocjene. Koristi se za poboljšanje performansi učenja s malo primjera, pronalaženje boljih „zero-shot“ lanaca razmišljanja, te usmjeravanje modela prema istinitosti i/ili informativnosti.

APE postiže razinu performansi koja je usporediva ili bolja u odnosu na ljudsku na upitima tipa *zero-shot* i odabranim zadacima iz standardnih mjerila kao što je npr. BIG-Bench. Automatski generirani upiti također mogu poboljšati performanse učenja s nekoliko primjera (*few-shot* *learning*) jednostavnim dodavanjem tih upita standardnim upitima za učenje iz konteksta. Ovaj pristup automatizira proces inženjerstva upita za LLM-ove predstavljajući ga kao problem optimizacije crne kutije [24].

Active Prompting

*Active Prompting* je metoda koja se koristi za poboljšanje performansi velikih jezičnih modela u složenim zadacima koji zahtijevaju razmišljanje, kao što su aritmetičko i zdravorazumsko rezoniranje. Cilj je prilagoditi LLM-ove specifičnim zadacima pomoću primjera zadataka s ljudski dizajniranim razmišljanjem u lancu misli. Problem je u tome što primjeri možda nisu najučinkovitiji primjeri za različite zadatke. Metoda identificira koja su pitanja najvažnija i najkorisnija iz skupa specifičnih upita za određeni zadatak.

Prvo se generira više odgovora za određeni skup pitanja kako bi se procijenila nesigurnost modela u vezi s tim pitanjima. Na temelju procijenjene nesigurnosti odabiru se najnesigurnija pitanja za koja se smatra da će biti najkorisnija. Zatim se piše lanac razmišljanja i odgovora za odabrana pitanja kako bi se stvorili novi primjeri. S novim primjerima, model se koristi za izvršavanje složenih zadataka razmišljanja. Ovaj pristup omogućuje LLM-u da se prilagodi specifičnim zadacima i poboljša svoje sposobnosti rješavanja problema kroz aktivno učenje i anotaciju [37].

Program-aided Language Models (PAL)

Program-aided Language Models je novi pristup koji koristi velike jezične modele za čitanje problema na prirodnom jeziku i generiranje programa slično kao CoT, ali rješavanje problema prepušta Python interpreteru. PAL koristi modele za dekompoziciju problema na izvršive korake, dok se samo rješavanje delegira interpreteru, što omogućuje preciznije rezultate u matematičkim, simboličkim i algoritamskim zadacima nego što to mogu veći modeli kao npr. PaLM-540B [38].

PAL postiže vrhunske rezultate na raznim zadacima razumijevanja prirodnog jezika, nadmašujući popularne tehnike postavljanja upita LLM-ovima. Ova tehnika koristi kombinaciju prirodnog jezika i programskih jezika, što omogućuje modelu da generira program koji će pružiti odgovor na zadano pitanje.

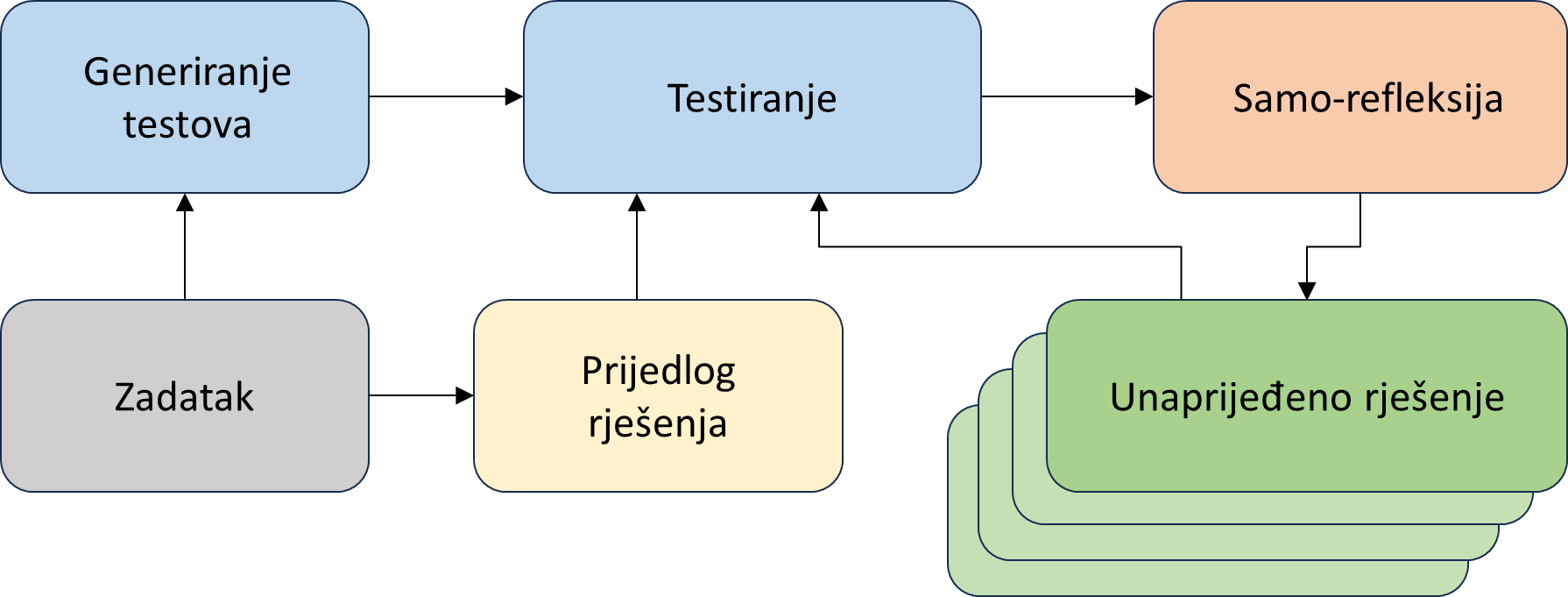
ReAct (Reasoning and Acting)

*ReAct* je pristup koji kombinira razmišljanje (npr. *CoT*) i djelovanje (npr. generiranje plana akcije) u modelima velikih jezičnih modela kako bi poboljšao njihovu sposobnost rješavanja zadataka i donošenja odluka. Razmišljanje pomaže modelu u indukciji, praćenju i ažuriranju planova akcije te u upravljanju iznimkama, dok djelovanje omogućuje interakciju s vanjskim izvorima za dodatne informacije. Omogućuje jezičnim modelima da generiraju tragove razmišljanja i specifične akcije za zadatak na međusobno povezan način, što dovodi do dinamičkog razmišljanja i prilagodbe planova djelovanja. Primjenjuje se na različite zadatke u području donošenja odluka, pokazujući bolju učinkovitost u odnosu na standardne metode i poboljšanu interpretativnost i pouzdanost [39].

Reflexion

*Reflexion* je tehnika postavljanja upita koja uključuje analizu vlastitih izlaza, ponašanja, znanja ili procesa zaključivanja od strane velikih jezičnih modela. Svrha ove tehnike je smanjenje „halucinacija“ trenutnih generativnih modela. Proces tehnike reflexion prikazuje Slika 3.10.

Arhitektura uključuje povratnu petlju, koja ima za cilj ispravljanje uobičajenih pogrešaka i može se smatrati pristupom *"LLM-in-the-loop"* umjesto pristupa *"man-in-the-loop“*. Temelji se na pretpostavci da model može djelovati i kao kreator i kao ocjenjivač svojih izlaza. Nakon obavljanja zadatka, model koristi povratne informacije koje je sam generirao kako bi poboljšao svoj početni odgovor [40].



Slika 3.10 Proces tehnike reflexion, po uzoru na [41]

Multimodal Chain-of-Thought Reasoning

Ovaj pristup kombinira jezične (tekst) i vizualne (slike) modalitete kao ulazne informacije u procesu zaključivanja. Koristi se za generiranje međukoraka u razmišljanju kao osnovi za generiranje odgovora na složena pitanja. Koristi se za poboljšanje sposobnosti složenog zaključivanja modela, posebno u znanstvenim i multimodalnim zadacima gdje je potrebno razumijevanje i tekstualnog i vizualnog konteksta.

Korištenje ove tehnike pokazalo se učinkovitim u smanjenju halucinacija (generiranje netočnih ili irelevantnih informacija) i ubrzavanju konvergencije modela, što dovodi do točnijeg zaključivanja i boljih rezultata [42].

## Usporedba velikih jezičnih modela

U ovom odjeljku opisani su neki od najčešće korištenih velikih jezičnih modela i način kako se u interakciji s njima uspješno primjenjuju najnovije i najnaprednije tehnike postavljanja upita. Osim toga, prikazane su mogućnosti ovih modela za obavljanje različitih zadataka i mogućnosti primjene nekih od tehnika postavljanja upita koje su opisane u prethodnim poglavljima. Razumijevanje ovih mogućnosti važno je za razumijevanje ograničenja velikih jezičnih modela i načina kako ih učinkovito koristiti.

### Mjerila za usporedbu velikih jezičnih modela

U području velikih jezičnih modela promjene su vrlo česte i mjerenje učinkovitosti modela ne može se svesti samo na prebrojavanje parametara i količine obrađenih podataka, već je potrebno uzeti u obzir „kognitivne“ sposobnosti modela te razumijevanje zadataka i konteksta. Stoga su definirana standardna mjerila koja služe kako bi se testirale sposobnosti jezičnih modela u raznim kategorijama, od osnovne matematike do složenog zaključivanja. Kako jezični modeli postaju sofisticiraniji, tako i kriteriji za mjerenje njihovih sposobnosti moraju postati napredniji.

U nastavku su opisana neka od trenutno najkorištenijih mjerila na temelju kojih je napravljena usporedba modela kasnije u poglavlju, a to su:

* MMLU
* HellaSwag
* AI2 Reasoning Challenge (ARC)
* WinoGrande
* HumanEval
* DROP
* GSM-8K
* MATH
* BIG-Bench-Hard
* Natural2Code
* WMT23
* GPQA (Diamond)
* MGSM

Massive Multitask Language Understanding (MMLU)

Razvijen od strane Facebook AI Research (FAIR), MMLU je opsežno mjerilo koje procjenjuje jezične modele kroz mnoštvo predmeta. Od prirodnih do društvenih znanosti, testira sustave umjetne inteligencije da pokažu svoje razumijevanje u područjima kao što su pravo, matematika i povijest. Cilj ovog mjerila je provjeriti koliko je model sposoban generalizirati svoje znanje i primijeniti ga u različitim kontekstima, pokrivajući kako osnovne tako i napredne razine pitanja. Rezultati na MMLU testu pružaju uvid u sveobuhvatne sposobnosti modela u jezičnom razumijevanju i rješavanju problema [43].

Grade School Math 8K (GSM8K)

Izrađen od strane stručnjaka Instituta Allen za AI, GSM8K je test za jezične modele u području osnovne matematike. Ovo mjerilo zahtijeva od umjetne inteligencije da se snalazi u osnovama aritmetike, algebre i geometrije, odražavajući izazove s kojima se susreću učenici osnovne škole, stoga procjenjuje praktičnu primjenu u obrazovnim okruženjima. Ovaj test ima oko osam tisuća zadataka iz različitih matematičkih koncepta kao što su osnovne operacije, geometrija, algebarski izrazi itd. Cilj testa GSM8K je procijeniti sposobnost modela da uspješno rješava osnovne matematičke probleme koji se obično javljaju u osnovnim školama [44].

Math Aptitude Test Heuristics (MATH)

MATH mjerila služe kao rigorozno testiranje numeričkih i logičkih sposobnosti umjetne inteligencije. Iako se tvorac ovog specifičnog mjerila može razlikovati, svrha je ista: testiranje granica sposobnosti računanja i logičkog zaključivanja sustava umjetne inteligencije. Ovaj test ima za cilj procijeniti kreativnost, brzinu razmišljanja i sposobnost rješavanja matematičkih problema kroz primjenu heurističkih metoda [44].

HumanEval

HumanEval je mjerilo definirano od strane tvrtke za evaluaciju sposobnosti sustava umjetne inteligencije u generiranju funkcionalnog programskog koda. Sastoji se od 164 programerska problema s jediničnim testovima (engl. *unit test*). Problemi služe kako bi se ocijenilo razumijevanje jezika, algoritama i jednostavne matematike. Također se ocjenjuje generiranje valjanih tijela funkcija na temelju zadanih potpisa funkcije i dokumentacije, postavljajući temelje za potencijal umjetne inteligencije u automatizaciji razvoja softvera. Cilj je ocijeniti sposobnost modela da generira točan kod koji zadovoljava zadane specifikacije i ispravno prolazi kroz sve testne slučajeve. Ovaj proces obično uključuje procjenu različitih aspekata modela, kao što su kvaliteta generiranih tekstova, točnost odgovora na postavljena pitanja ili razumijevanje konteksta [45].

Beyond the Imitation Game Benchmark - Hard (BIG-Bench-Hard)

BIG-Bench mjerilo za procjenu sposobnosti jezičnih modela osmišljeno je kako bi se dobile korisne informacije za buduća istraživanja, pripremio teren za nove sposobnosti modela i ublažilo društveno štetne učinke. BIG-bench obuhvaća 204 zadatka iz raznih područja kao što su lingvistika, matematika, razvoj djeteta, biologija, fizika i mnoge druge discipline. Evaluacija se provodi procjenom ponašanja jezičnih modela tijekom izvršavanja zadataka, a rezultati uključuju usporedbu performansi ovih modela s ljudskim ocjenjivačima, čime se pruža čvrsta usporedna osnova za daljnje analize i unapređenja [46].

Discrete Reasoning Over Paragraphs (DROP)

Ovo mjerilo, koje je razvio Institut Allen, zapravo je skup podataka i testni set za procjenu sposobnosti modela da razumije i rješava složene jezičke zadatke koji zahtijevaju diskretno zaključivanje nad paragrafima teksta. DROP je dizajniran kako bi potaknuo istraživanje metoda koje kombiniraju distribuirane reprezentacije s diskretnim rezoniranjem. Sustavi moraju razumjeti odlomak teksta dovoljno dobro da odgovore na proizvoljna pitanja o njemu. Ovaj test se fokusira na različite aspekte jezičnog razumijevanja, uključujući analizu, zaključivanje i odgovaranje na pitanja na osnovu informacija danih u tekstu. Sadrži raznolike zadatke koji zahtijevaju različite razine razmišljanja, od osnovnih do naprednih, čime pruža sveobuhvatan pregled sposobnosti modela u diskretnom zaključivanju nad tekstualnim podacima [47].

HellaSwag

Ovo mjerilo, koje je također razvio Institut Allen za AI, koristi se za kvantificiranje zdravorazumskog zaključivanja. To je izazovan skup pitanja s višestrukim izborom koji ocjenjuje prediktivne sposobnosti modela u stvarnim scenarijima testirajući granice njegovog razumijevanja prirodnog jezika. Ovaj test se fokusira na razumijevanje konteksta i sposobnost donošenja zaključaka na osnovu suptilnih ili neintuitivnih informacija u tekstu, što ga čini izazovnim za modele koji se oslanjaju samo na površinsku analizu jezičnih podataka [44].

AI2 Reasoning Challenge (ARC)

Ovo mjerilo sastoji se od seta testova i skupa podataka koji su osmišljeni kako bi se istraživale sposobnosti modela umjetne inteligencije u različitim vrstama logičkog i deduktivnog zaključivanja. ARC za cilj ima promicanje istraživanja u naprednom odgovaranju na pitanja, posebno pitanja koja zahtijevaju rasuđivanje, korištenje zdravog razuma i drugih metoda za dublje razumijevanje teksta. Konkretno, pitanja ARC Challenge-a su ona na koja je teško odgovoriti jednostavnim osnovnim crtama. ARC izazov uključuje raznovrsne zadatke iz područja tehničkih znanosti, matematike, književnosti i društvenih znanosti, koji zahtijevaju duboko razumijevanje teksta i sposobnost logičkog zaključivanja [48].

Winogrande

Ovaj veliki skup podataka za testiranje nadahnut je izvornim WSC testom osmišljenim za procjenu sposobnosti strojeva u razumijevanju prirodnog jezika i zdravorazumskog zaključivanja, ali prilagođen kako bi se poboljšao opseg i težine skupa podataka. Najbolji modeli na ovom skupu podataka dosežu točnost 15 - 35% lošiju od ljudske izvedbe, ovisno o količini dostupnih podataka za treniranje. Cilj ovog mjerila je istražiti prave sposobnosti modela za zdravorazumsko zaključivanje i smanjiti pristranost u podacima koja može dovesti do preuveličavanja stvarnih sposobnosti velikih jezičnih modela [49].

Natural2Code

Natural2Code je metoda za evaluaciju sposobnosti modela za generiranja koda, koji koriste prirodni jezik kao ulaz i programski kod kao izlaz. Zadatak generiranja koda iz prirodnog jezika smatra se zahtjevnim izazovom. Ocjenjuje se sposobnost modela u korištenju naprednih algoritama obrade prirodnog jezika kako bi se upiti napisani prirodnim jezikom interpretirali u odgovarajući programski kod. LLM-ovi postižu impresivne rezultate u NL2Code zadatku, ali i dalje postoje izazovi i prilike za poboljšanje [50].

WMT23

WMT je standardizirano mjerenje kvalitete strojnog prevođenja s jednog prirodnog jezika na drugi na godišnjem natjecanju za računalni prijevod koje organizira Konferencija o strojnom prevođenju (Conference on Machine Translation). WMT23 se sastoji od nekoliko metrika kojima se mjeri kvaliteta strojnih prijevoda na raznim jezičnim parovima u odnosu na profesionalne ljudske prijevode. Glavne metrike su [51]:

* BLEU - mjeri preciznost prevođenja u odnosu na referentni prijevod
* TER - mjeri broj potrebnih uređivanja kako bi se automatizirani prijevod transformirao u referentni
* ChrF - slična BLEU-u, ali osjetljivija na sinonime i poredak riječi
* BLEURT (Bilingual Evaluation Understudy Reference-less Training) - novija metrika za procjenu kvalitete strojnih prijevoda razvijena od strane Google AI tima. Za razliku od BLEU-a i drugih starijih metrika koje zahtijevaju referentne ljudske prijevode za evaluaciju, BLEURT je posebno dizajniran da ne zahtijeva ljudske referentne prijevode. Koristi pristup dubokog učenja gdje se model obučava na ogromnom skupu sintetičkih i ljudski procijenjenih prijevoda da nauči procijeniti kvalitetu prijevoda bez potrebe za referentnim prijevodom.

Ove automatske metrike se temelje na usporedbi sa zlatnim standardom ljudskih prijevoda za razne tekstualne domene kao što su novinski članci, web stranice, govori itd. Služe za rangiranje različitih sustava za strojno prevođenje prema njihovoj učinkovitosti.

GPQA

GPQA (Graduate-Level Google-Proof Q&A Benchmark) je izazovan skup podataka s pitanjima višestrukog izbora koja su napisali stručnjaci u području biologije, fizike i kemije. Skup podataka služi za testiranje sposobnosti sustava UI da odgovaraju na vrlo teška pitanja, s ciljem razvoja metoda nadzora koje omogućuju ljudima da nadgledaju izlaze sustava UI. Postoje tri verzije testnog skupa s različitim brojem pitanja: prošireni (546), glavni (448) i dijamant (198). GPQA je važno mjerilo za ocjenjivanje LLM-ova, s obzirom na hitnu potrebu za modelima koji mogu obraditi i generirati točne informacije u specijaliziranim znanstvenim područjima [52].

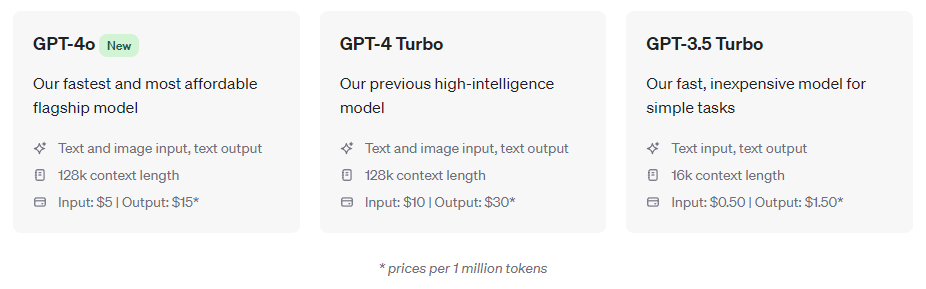
MGSM

MGSM (Multilingual Grade School Math) je višejezični matematički test koji ocjenjuje sposobnost rješavanja aritmetičkih problema kod velikih jezičnih modela. Sadrži 250 matematičkih problema iz osnovnoškolske matematike prevedenih na deset različitih jezika. Koristi se za procjenu sposobnosti višejezičnog razmišljanja kod velikih jezičnih modela, posebno kada se koristi metoda lanca misli (CoT) [53].

U sljedećem dijelu opisani su neki od trenutno najnaprednijih velikih jezičnih modela (GPT-4, Claude 3 Opus, Gemini 1.5 Pro i LLama 3), neke specifičnosti interakcije s njima, te je dana usporedba performansi modela u odnosu na neka standardna mjerila.

### GPT-4

Tvrtka OpenAI[[1]](#footnote-2) razvila je nekoliko modele za generiranje teksta koji se nazivaju generativni predtrenirani transformeri ili skraćeno "GPT" modeli. Takvi modeli poput GPT-4 i GPT-3.5, trenirani su za razumijevanje prirodnog i formalnog jezika. Modeli poput GPT-4 mogu se koristiti za veliki broj različitih tipova zadataka uključujući generiranje tekstualnog sadržaja ili programskog koda, sažimanje teksta, razgovor, kreativno pisanje, itd. Slika sadrži kratak opis tri trenutno najpopularnija GPT modela.



Slika 3.11Trenutno najnapredniji GPT modeli

ChatGPT je LLM model obučen od strane tvrtke OpenAI koji ima sposobnost interakcije u obliku razgovora. Ovaj model je osposobljen slijediti upute u upitu kako bi pružio odgovarajuće odgovore u kontekstu dijaloga. ChatGPT može pomoći u odgovaranju na pitanja, predlaganju recepata, pisanju tekstova u određenom stilu, generiranju koda i još mnogo toga. ChatGPT se obučava korištenjem pojačanog učenja iz ljudskih povratnih informacija (engl. *Reinforcement learning from Human Feedback*, RLHF). Iako je ovaj model puno sposobniji od prethodnih iteracija GPT-a, a također i obučen za smanjenje štetnih i neistinitih rezultata, još uvijek ima određena ograničenja.

U prethodnim poglavljima su opisani neki od načina kako usmjeriti LLM da vodi razgovor u određenom stilu, s određenom namjerom, ponašanjem i identitetom. Ovdje možemo vidjeti jedan osnovni primjer interakcije sa ChatGPT modelom koji predstavlja sustav za razgovor s mogućnošću generiranja odgovora na više tehnički i znanstveni način [14].

*„U nastavku slijedi razgovor s asistentom za istraživanje koji se temelji na umjetnoj inteligenciji. Ton asistenta je tehnički i znanstveni.*

*Korisnik: Pozdrav, tko si ti?*

*AI asistent: Pozdrav! Ja sam asistent za istraživanje koji se temelji na umjetnoj inteligenciji. Kako vam mogu pomoći danas?*

*Korisnik: Možeš li mi reći nešto o stvaranju crnih rupa?*

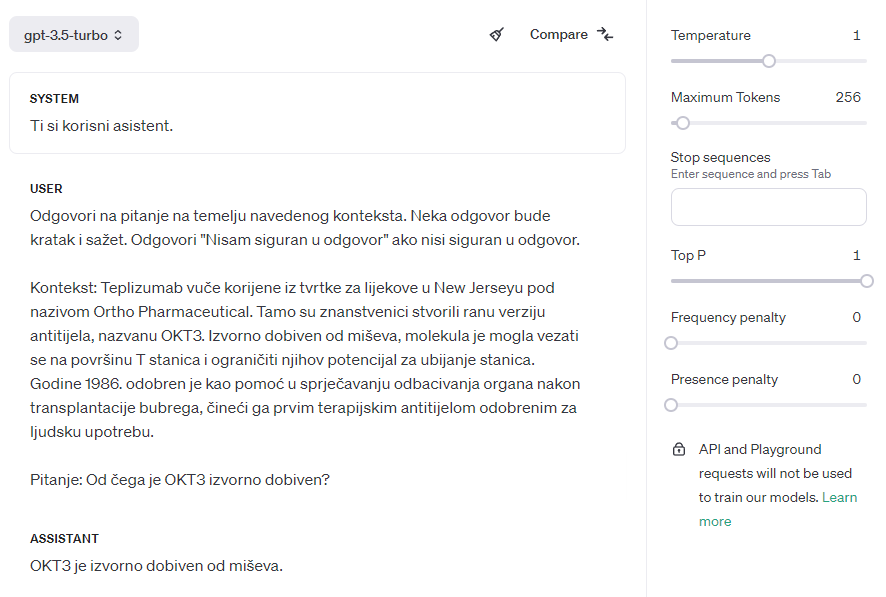
*AI asistent: Naravno! Crne rupe nastaju kada masivna zvijezda ostane bez goriva i kolabira pod vlastitom gravitacijom. Ovaj kolaps uzrokuje da zvijezda postane nevjerojatno gusta, s gravitacijskim poljem toliko jakim da joj čak ni svjetlost ne može pobjeći. …“*

U ovom primjeru mogu se uočiti dvije važne komponente:

* namjeru ili objašnjenje što je *chatbot*
* identitet koji upućuje na stil ili ton koji će chatbot koristiti za odgovor

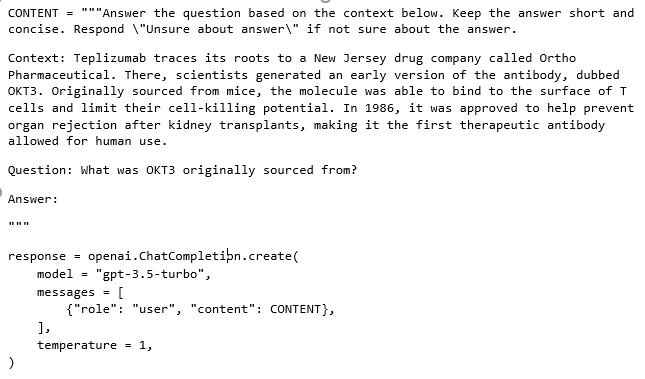
Format razgovora omogućuje višestruko postavljanje upita i dobivanje odgovora u nekoliko koraka, ali podržava i jednokratne zadatke u kojima se pokušava dobiti odgovarajući rezultat pomoću samo jednog upita kao što se moglo vidjeti u primjeru „Odgovaranje na pitanja“ u poglavlju 3.2.1.

Sljedeća slika pokazuje kako se isti zadatak može izvršiti koristeći *OpenAI Playground* okruženje u kojem se osim opisa konteksta, određivanja uloge i davanja instrukcija kroz upit, interakcija može oblikovati i konfiguracijom postavki modela koje su objašnjene u poglavlju 3.1.



Slika 3.12 OpenAI Playground okruženje

Interakcija s ChatGPT modelom može se ostvariti i tako da se koristi API preko kojega se šalje upit koji sadrži sve postavke i elemente upita kao što su npr. temperatura i kontekst, kao što je vidljivo na primjeru kojeg prikazuje Slika 3.13. Glavni unos je parametar poruka. Poruke moraju biti niz objekata poruka, gdje svaki objekt ima ulogu ("sustav", "korisnik" ili "asistent") i sadržaj. Kao što je spomenuto, ti razgovori mogu biti kratki, sa samo jednom porukom, ili se sastojati od više uzajamnih razmjena. Poruka sustava pomaže postaviti ponašanje asistenta. Na primjer, može se prilagoditi osobnost asistenta ili dati specifične upute o tome kako bi se trebao ponašati tijekom razgovora. Poruke korisnika pružaju zahtjeve ili komentare na koje asistent odgovara. Prethodni odgovori asistenta se pohranjuju i oblikuju kontekst, ali mogu se i unaprijed napisati u upitu kao primjeri željenog ponašanja. U slučaju kada korisničke upute upućuju na prethodne poruke važno je uključiti povijest razgovora [14].



Slika 3.13 Primjer oblikovanja upita za Open API

Možemo vidjeti da se radi o modelu gpt-3.5-turbo, da je definiran kontekst, temperatura te instrukcije da odgovor treba biti koncizan i da se ne daje odgovor ukoliko nije siguran. Na sličan način možemo dohvatiti odgovor od modela. Primjer odgovora prikazuje Slika 3.14.



Slika 3.14 Primjer odgovora kojeg šalje API

Svaki odgovor od API-ja uključuje parametar *finish\_reason* koji označava zašto je generativni proces završen*,* a moguće vrijednosti tog parametra su:

* *stop*: API je vratio potpunu poruku ili poruku prekinutu jednom od sekvenci zaustavljanja navedenih putem parametra zaustavljanja
* *length*: Nepotpun izlaz modela zbog parametra *max\_tokens* ili ograničenja tokena
* *function\_call*: Model je odlučio pozvati funkciju
* *content\_filter*: izostavljeni sadržaj zbog oznake naših filtara sadržaja
* *null*: odgovor API-ja još uvijek je u tijeku ili nije potpun

Ovisno o ulaznim parametrima, odgovor modela može uključivati ​​različite informacije. Jedan način korištenja razgovora je navođenje modela da vrati objekt u JSON formatu.

GPT-4

GPT-4 je veliki multimodalni model koji može obrađivati unose koji uključuju kombinaciju slika i teksta, generirajući tekstualne izlaze na temelju kombiniranih informacija. Može obrađivati upite koji sadržavaju tekst i slike, omogućujući širok raspon zadataka poput analize dokumenata, opisa fotografija itd. Model može opisivati i analizirati slike, što je prikazano u tehničkom izvješću modela s primjerom duhovitog opisa slike. Upit se sastoji od pitanja o slici (Slika 3.15) na koje GPT-4 može odgovoriti [54].

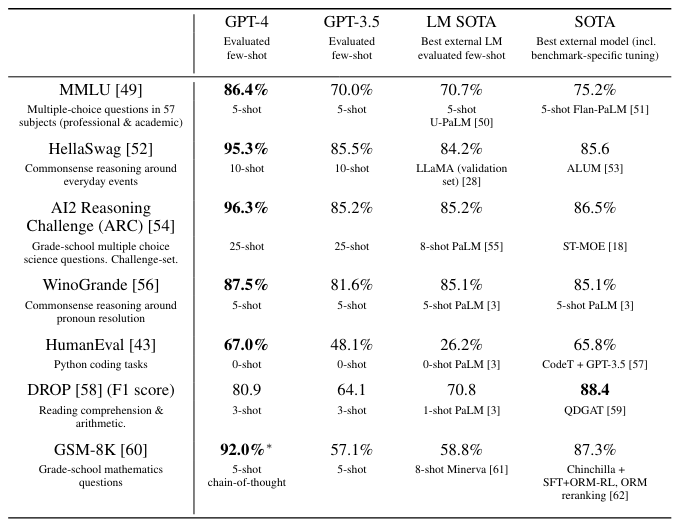


Slika 3.15 Primjer upita koji demonstrira vizualne mogućnosti unosa modela GPT-4 [54].

Ovako izgleda tekstualni dio upita i odgovor modela:

* *Korisnik: Što je smiješno na ovoj slici? Opisuj je panel po panel.*
* *GPT-4: Slika prikazuje paket za adapter "Lightning Cable" s tri panela.*
  + *Panel 1: Pametni telefon s VGA konektorom (velikim, plavim, 15-pinskim konektorom koji se obično koristi za računalne monitore) priključenim u njegovu priključnu luku za punjenje.*
  + *Panel 2: Paket za adapter "Lightning Cable" s slikom VGA konektora na njemu.*
  + *Panel 3: Krupni plan VGA konektora s malim Lightning konektorom (koji se koristi za punjenje iPhonea i drugih Apple uređaja) na kraju.*
  + *Humor na ovoj slici proizlazi iz apsurdnosti priključivanja velikog, zastarjelog VGA konektora u malu, moderni utor za punjenje pametnog telefona.*

Model GPT-4 potiže razinu performansi sličnu ljudskoj na raznim referentnim testovima (Slika 3.16), ostvarujući rezultate u top 10% na simuliranom ispitu za pravosudni ispit. U usporedbi s prethodnim modelima, GPT-4 smanjuje “halucinacije” i greške u rezoniranju, što ga čini pouzdanijim. Unatoč napretku, GPT-4 još uvijek ima ograničenja i predstavlja nove sigurnosne izazove zbog svojih poboljšanih sposobnosti [54].

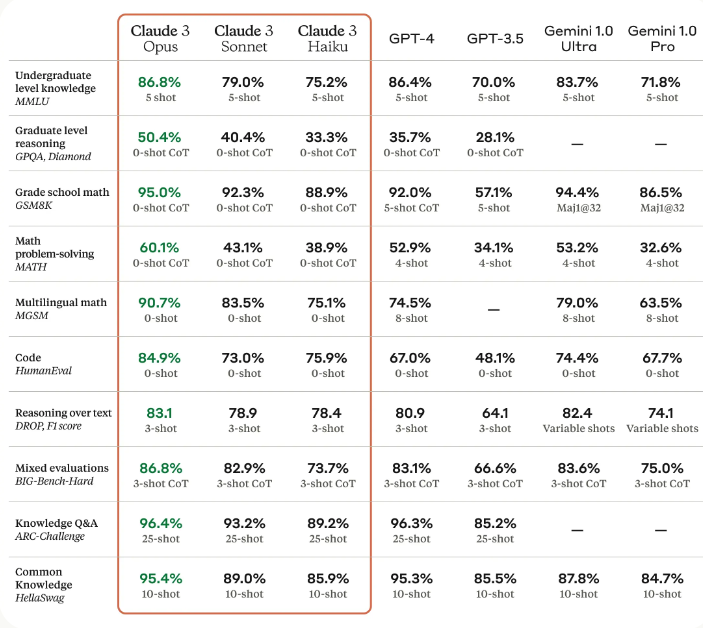


Slika 3.16 Performanse modela GPT-4 prema standardnim mjerilima

### Claude 3 Opus

Claude je obitelj velikih jezičnih modela razvijenih od strane tvrtke Anthropic[[2]](#footnote-3) čiji su osnivači bivši djelatnici tvrtke OpenAI. Claude 3 obitelj modela uključuje tri najsuvremenija modela poredana uzlazno prema mogućnostima: Claude 3 Haiku, Claude 3 Sonnet i Claude 3 Opus. Svaki sljedeći model nudi sve moćnije performanse, dopuštajući korisnicima odabir optimalne ravnoteže inteligencije, brzine i cijene za njihovu specifičnu primjenu. Svi modeli Claude 3 obitelji imaju mogućnosti multimodalnog unosa s tekstualnim izlazom. Modeli mogu obrađivati slikovne podatke i imaju poboljšanu fluentnost u ne-engleskim jezicima, što ih čini svestranima za globalnu publiku. Dizajnirani su da pomažu u zadacima kao što su kodiranje, pisanje, uređivanje i sažimanje, te mogu interpretirati vizualne unose. Slijede konverzacijski kontekst i mogu se prilagoditi različitim "osobnostima" [55].

Claude 3 modeli pokazuju snažne performanse na referentnim testovima za rasuđivanje, matematiku i kodiranje. Kao što prikazuje Slika 3.17., najbolja verzija ovog modela nadmašuje GPT-4 i Gemini 1.0 modele na standardnim mjerilima kao što su MMLU i HumanEval.



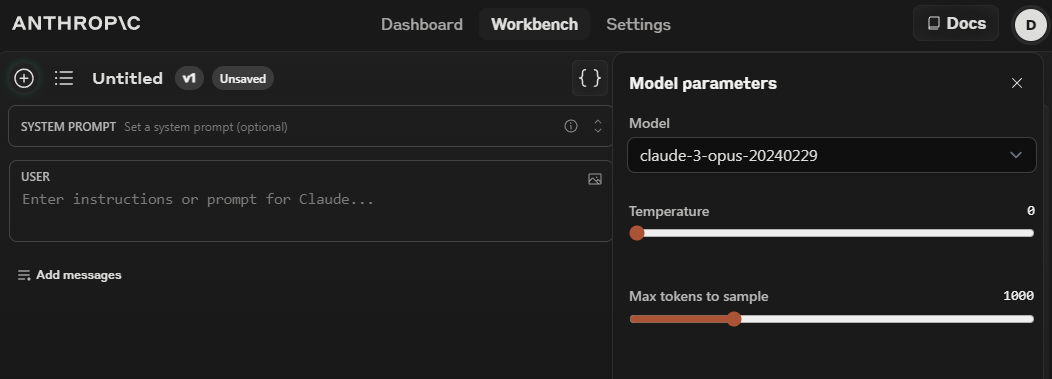
Slika 3.17 Performanse modela Claude 3 u odnosu na GPT i Gemini modele

Claude 3 Opus pokazuje gotovo dvostruko veću točnost na određenim testovima usporedbi s Claude 2.1 modelom, uz značajno smanjenje netočnih odgovora. Unatoč poboljšanjima, postoji prostor za optimizaciju kako bi se netočni odgovori preusmjerili na 'Ne znam/Nisam siguran' odgovore. Kontekstualni prozor je proširen čime su se dobile nove mogućnosti poput analize više dokumenata.

Opus također pokazuje značajna poboljšanja u odgovaranju na činjenična pitanja u otvorenim oblicima pitanjima, dok smanjuje netočne odgovore ili halucinacije. Modeli Claude 3 također su bolji od prethodnih u stvaranju strukturiranih izlaza poput JSON objekata.

Iako Claude modeli pokazuju značajan napredak u određenim sposobnostima, povremeno mogu generirati netočne ili zavaravajuće informacije. Kao i drugi LLM-ovi, Claude može proizvesti uvredljiv ili pristran sadržaj. Važno je imati na umu da Claude nije namijenjen davanju savjeta, uključujući pravne, financijske i medicinske savjete. Stoga se ne bi trebalo oslanjati isključivo na odgovore velikih jezičnih modela bez provođenja vlastitih neovisnih istraživanja.

Modeli Opus i Sonnet su dostupni za korištenje kroz proizvode claude.ai, Anthropic API i okruženje Anthropic Console, koje prikazuje Slika 3.18, što je pandan OpenAI playground okruženju. Također su dostupne biblioteke u Pythonu i Typescriptu kako bi se olakšao rad s Anthropic API-jem i postoji mogućnost validacije upita kako bi se osiguralo da su pravilno formatirani za API.



Slika 3.18 Anthropic Console okruženje

Unutar okruženja *Anthropic* *Console* dostupne su postavke *temperature* i *max* *tokens*, međutim ukoliko se koristi API moguće je konfigurirati i ostale postavke opisne u poglavlju 3.1 kao što su St*op Sequence* i *Top P* ali i dodatne postavke poput *metadata*, *system* i *tools* [56].

### Gemini 1.5 Pro

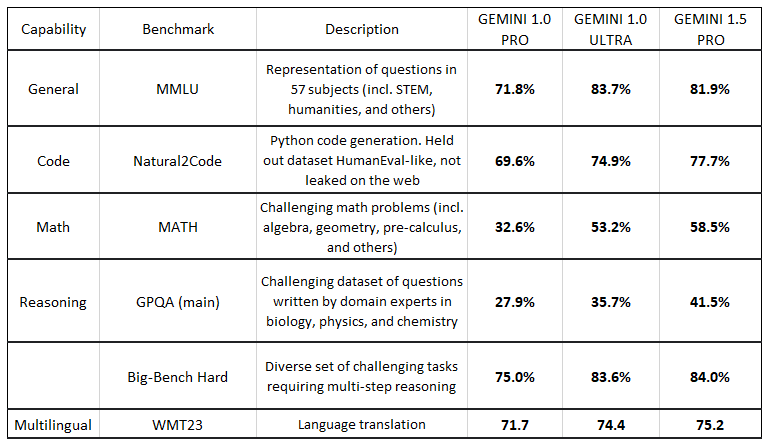
Gemini 1.5 Pro je najnoviji i najsposobniji model UI razvijen od strane Google Deepmind-a. Podržava multimodalnu interakciju i pokazuje impresivnu sposobnost zaključivanja na temelju ulaznih podataka u obliku tekstova, slika, audio i video zapisa te programskog koda.

Gemini modeli dolaze u tri veličine:

* Ultra - najnapredniji model serije, dobar za vrlo složene zadatke
* Pro - smatra se najboljim modelom za skaliranje kroz širok raspon zadataka
* Nano - učinkovit model za zadatke i slučajeve korištenja s ograničenom memorijom na uređaju; uključuje modele sa 1.8 milijardi (Nano-1) i 3.25 milijardi (Nano-2) parametara

Gemini 1.5 modeli donose značajna unaprjeđenja u performansama, posebno u obradi dugih konteksta. Verzija 1.5 Pro nadmašuje prethodne verzije Gemini modela i brojne druge modele u zadacima s dugim kontekstom. Njegove mogućnosti obrade dugih konteksta omogućuju mu da procesuira cijele kolekcije dokumenata, sate video materijala i dane zvučnih zapisa [57].

Prema tehničkom izvješću, Gemini nadmašuje najsuvremenije modele s kojima se usporedio u velikoj većini kategorija usporedbe poput obrade jezika, kodiranja, zaključivanja i multimodalnog zaključivanja. Slika 3.19 pokazuje usporedbu verzije 1.5 Pro s prethodnim verzijama u nekim kategorijama.



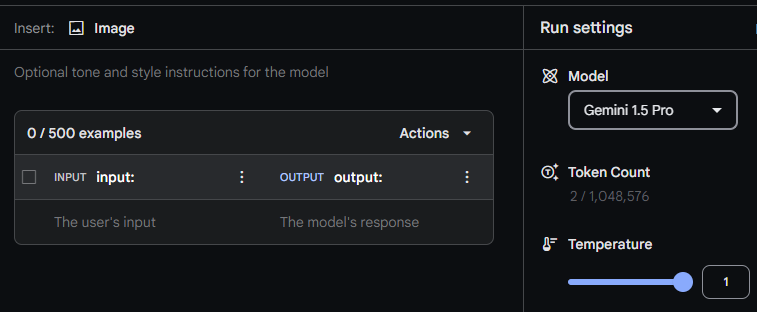
Slika 3.19 Performanse Gemini 1.5 Pro modela [57]

Model Gemini se smatra prvim modelom koji postiže performanse na razini ljudskog stručnjaka u odnosu na MMLU ispitno mjerilo i postaje novi predvodnik u 20 multimodalnih mjerila. Najnoviji Gemini modeli imaju sposobnost obrade milijuna tokena što otvara praktične primjene koje prije nisu bile moguće. Kao što prikazuje Slika 3.20, Gemini 1.5 Pro može obraditi velike baze koda u cijelosti, poput JAX-a (746.152 tokena), i odgovoriti na vrlo specifične upite o njima.



Slika 3.20 Upotreba baze koda za kontekst upita [57]

Za korištenje Gemini modela također se može koristiti Gemini API kao i okruženje Google AI studio. Zanimljivost koju nudi ovo okruženje je opcija „*Structured prompt“* koja omogućuje kombiniranje instrukcija s primjerima kako bi se modelu pokazala preferirana vrsta izlaza, umjesto da mu se samo napiše što da radi. Slika 3.21 prikazuje ovu opciju unutar AI studio okruženja.



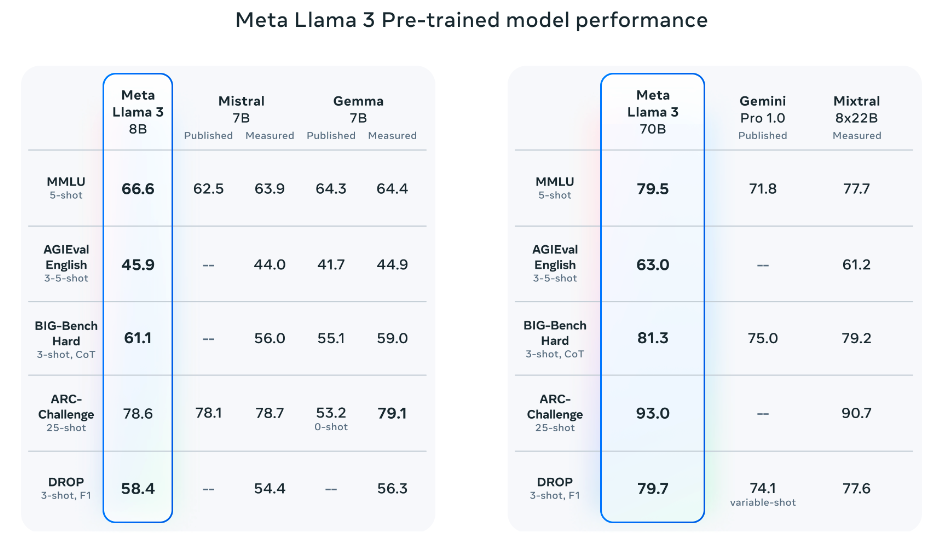
Slika 3.21 Structured prompt opcija

Ova tehnika postavljanja upita (*few-shot prompting*), korisna je kada želimo da model slijedi konzistentan format izlaza (npr. strukturirani JSON) ili kada je teško opisati riječima što želimo da model učini (npr. pisanje u određenom stilu). Primjeri se mogu unijeti ručno ili uvesti iz zasebne datoteke.

### LLama 3

Llama je obitelj velikih jezičnih modela razvijenih od strane tvrtke Meta[[3]](#footnote-4). U ovom odjeljku je opisana nova skupina Llama modela pod nazivom Llama 3, koji dolaze u dvije verzije (engl. *pre-trained, instruction-tuned*) i dvije veličine od 8 i 70 milijardi parametara.

Slika 3.22 prikazuje usporedbu s nekoliko drugih modela u odnosu na mjerila definirana na početku poglavlja. U usporedbi s modelima prikazanim na slici, Llama 3 „Pre-trained“ verzija modela postiže bolje rezultate, međutim u odnosu na ostale modele opisane u ovom poglavlju rezultati su ipak lošiji što se može vidjeti iz tablice (Slika 3.23) i grafa (Slika 3.24) koji objedinjuju rezultate sva četiri modela odabrana za usporedbu.



Slika 3.22 Performanse LLama 3 modela

Uz spomenute Llama modele postoji i Code Llama, obitelj velikih jezičnih modela, s mogućnostima prihvaćanja tekstualnih upita te generiranja programskog koda i rasprave o kodu. Postoje dvije varijante (Code Llama Python i Code Llama Instruct) te različite veličine (7B, 13B, 34B i 70B). Ovi modeli pokazuju dobre rezultate u generiranju i analiziranju python koda, pisanju unit testa i generiranju SQL upita.

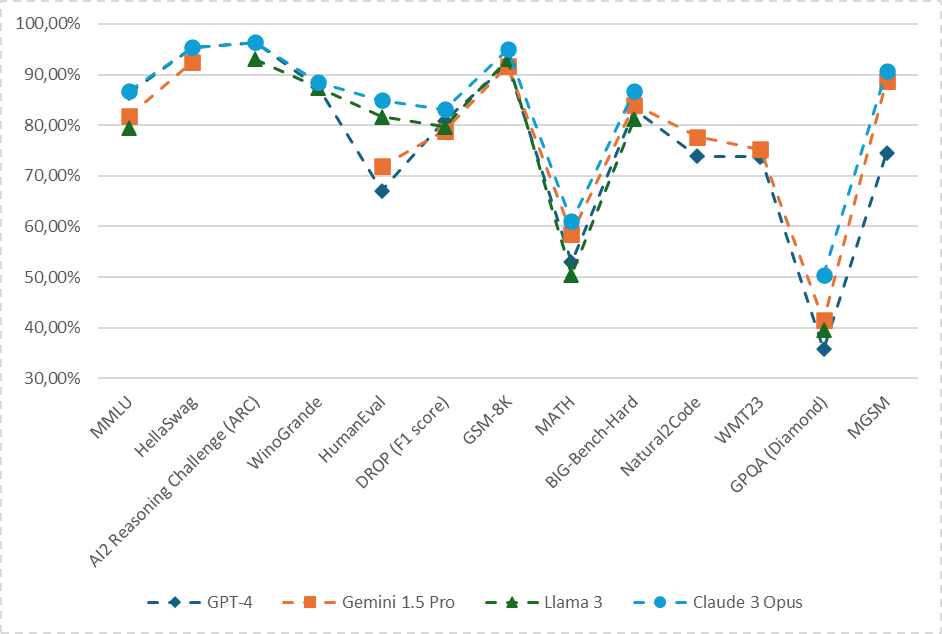
Tvrtka Meta je također izvijestila isporuku modela s 400 milijardi parametara koji se još uvijek usavršava i uskoro će biti dostupan. U pripremi je i podrška za multimodalnost, višejezičnost i dulje kontekstne prozora. Posljednji objavljeni podaci (od 15. travnja 2024.) za Llama 3 400B daju bolje rezultate na uobičajenim mjerilima kao što su MMLU 86.1% i Big-Bench Hard 85.3% čime se Llama približava trenutno najboljim rezultatima.

A screenshot of a table

Description automatically generated

Slika 3.23 Usporedba performansi modela prema standardnim mjerilima

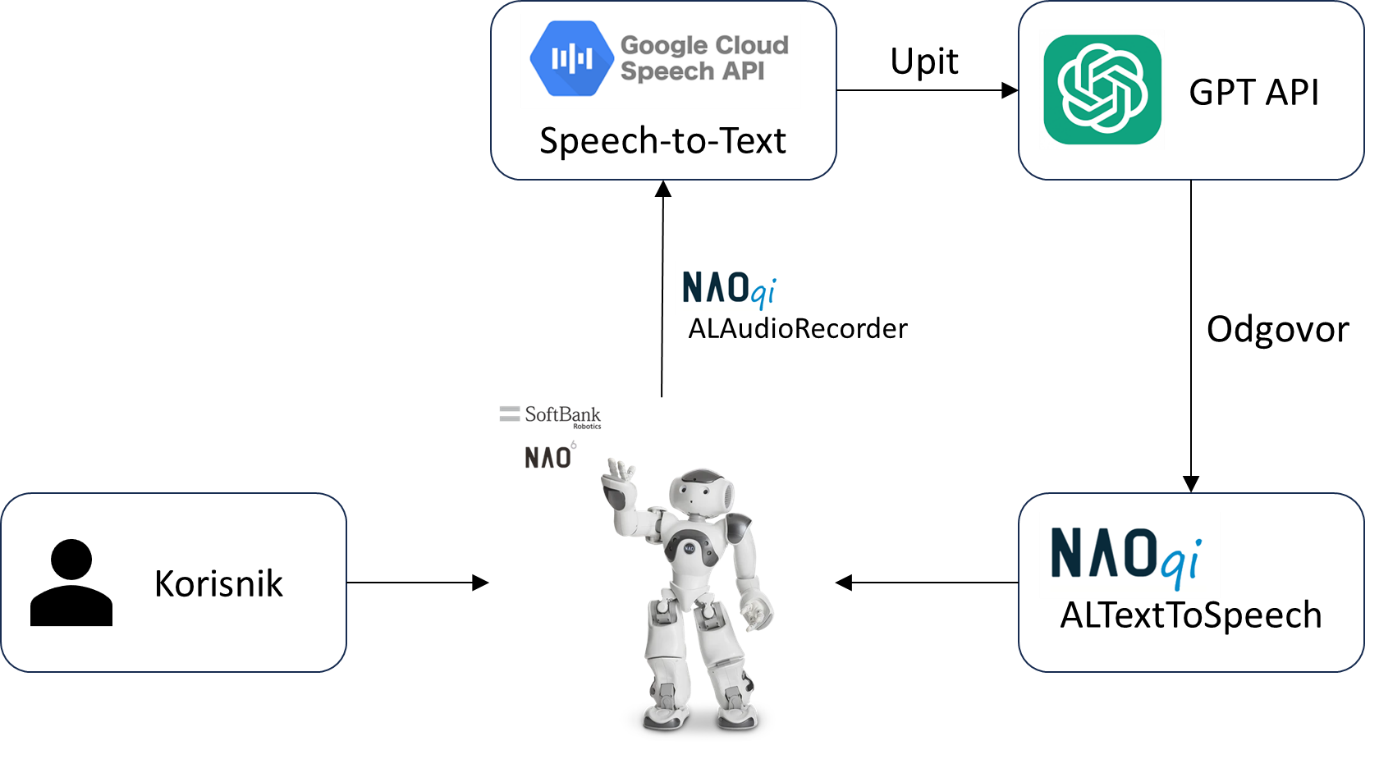
Svi podaci izvučeni su iz službene dokumentacije za svaki model, osim podataka u donjem dijelu stupca za GPT-4 model (pozadina u sivoj boji) koji su pronađeni izvan službene dokumentacije.



Slika 3.24 Grafički prikaz performansi modela

## Primjer korištenja upita za velike jezične modele: Interakcija s NAO robotom

Ovaj primjer predstavlja napredni oblik interakcije između čovjeka i robota, gdje se ljudski govor pretvara u tekstualni format kako bi se omogućila komunikacija s umjetnom inteligencijom. Slika 3.25 prikazuje sustav u kojem najprije korisnik s robotom komunicira govorom koji se detektira mikrofonom na robotu. Govor se zatim pretvara u tekst pomoću usluge Google Cloud Speech API, a tekst koji proizlazi iz tog procesa koristi se za stvaranje upita koji se šalje GPT modelu putem API-a. Model obrađuje upit i vraća odgovor, koji se potom šalje natrag robotu u obliku teksta. Robot zatim koristi modul NAOqi ALTextToSpeech za pretvaranje teksta u govor, čime se zatvara komunikacijski krug i omogućava robotu da verbalno odgovori korisniku.



Slika 3.25 Sustav za interakciju s NAO robotom

Ovaj sustav predstavlja primjer kako se tehnologija može koristiti za stvaranje interaktivnog iskustva gdje roboti mogu efikasno razumjeti i odgovarati na ljudske upite, pružajući osjećaj prirodne konverzacije.

Za uspješnu interakciju s robotom, ključno je definirati kontekstualne informacije koje omogućuju modelu da razumije situaciju u kojoj se nalazi. Ovo uključuje pružanje osnovnih informacija o samom robotu i njegovim sposobnostima, kao i specifične detalje vezane uz sami dijalog. Temeljne informacije obuhvaćaju ime, lokaciju i ulogu robota. Specifični detalji o dijalogu omogućuju modelu da prilagodi svoje odgovore u skladu s očekivanjima i potrebama korisnika, kao npr. duljinu odgovora.

Kako bi se postigao prirodniji i relevantniji dijalog, korištene su tehnike postavljanja upita. Ove tehnike uključivale su korištenje ključnih riječi i fraza koje su potaknule model da generira smislene i kontekstualno prikladne odgovore. Ovaj pristup je omogućio finu prilagodbu načina na koji robot interpretira upite i odgovara na njih, čime je poboljšana kvaliteta i relevantnost interakcije. Ovaj proces zahtijevao je pažljivo osmišljavanje i testiranje upita kako bi se osigurala njihova učinkovitost u različitim scenarijima razgovora.

Kako bi se osigurala efikasnost razgovora, ograničena je duljina odgovora modela. Skraćivanje odgovora omogućilo je brže razmjene informacija i smanjilo mogućnost da robot izgubi fokus tijekom dijaloga. Korištenjem konciznih i sažetih odgovora, poboljšana je dinamika interakcije, što korisnicima pruža ugodnije iskustvo.

Nakon implementacije inicijalnih postavki, proveden je niz testova kako bi se procijenila kvaliteta dijaloga i identificirala područja za poboljšanje. Testiranje je uključivalo simulacije stvarnih razgovora, analizu odgovora robota i prikupljanje povratnih informacija od korisnika. Na temelju tih povratnih informacija, kontinuirano se prilagođavao kontekst i tehnike postavljanja upita kako bi se poboljšalo interakciju.

# INTERAKCIJA SA SUSTAVIMA UMJETNE INTELIGENCIJE

Interakcija sa sustavima umjetne inteligencije postaje sve važnija tema kako se ti sustavi sve više integriraju u različite aspekte svakodnevnog života. Znanstvena disciplina koja se bavi istraživanjem interakcije ljudi i sustava UI obuhvaća proučavanje i osmišljavanje načina na koje ljudi komuniciraju i surađuju sa sustavima UI s ciljem stvaranja takvih sustava koji bi bili korisni ljudima, jednostavni za upotrebu, pouzdani i etični. Dakle, dizajn interakcije između ljudi i sustava UI trebao bi se temeljiti na principima koji osiguravaju sigurnost, razumljivost i učinkovitost, a osim toga potrebno je spriječiti neželjena ponašanja sustava te smanjiti rizike od nepredvidljivih i potencijalno opasnih posljedica korištenja umjetne inteligencije [58].

Jedan od glavnih izazova u oblikovanju UI sustava je adekvatno predstavljanje različitih korisničkih perspektiva. Naime, korisnici mogu imati različito predznanje, potrebe i očekivanja, što zahtijeva prilagodljive i personalizirane sustave UI koji se mogu prilagoditi potrebama, preferencijama, ciljevima i povratnim informacijama korisnika. U tu svrhu, istraživači područja interakcije čovjeka i računala (engl. *Human-Computer Interaction, HCI*) ističu potrebu za uključivanjem korisnika u proces dizajna i razvoja kako bi se osiguralo da sustavi zadovoljavaju njihove stvarne potrebe [58], [59].

Prema nekim autorima postoji devet dimenzija interakcije između čovjeka i sustava umjetne inteligencije [60]:

* Transparentnost: razina svjesnosti s kojom osoba komunicira sa sustavom UI, varirajući od potpune svjesnosti do potpune nesvjesnosti
* Impuls interakcije: početak i razlog nove interakcije između čovjeka i UI, koji može biti ciljani, istraživački ili kreativni
* Rezultat interakcije: kako ishod interakcije utječe na okolinu, s mogućnostima praćenja, informiranja, asistiranja, savjetovanja ili pružanja iskustva
* Smjer akcije: jednosmjerno od čovjeka prema sustavu UI ili obratno, ili dvosmjerno.
* Kanal akcije: senzorni oblik percepcije kroz koji se informacije prenose
* Frekvencija akcije: broj akcija unutar jedne interakcije
* Frekvencija interakcije: broj interakcija unutar određenog vremenskog razdoblja
* Ovisnost interakcije: razina ovisnosti između više interakcija
* Okolina interakcije: okolina u kojoj se interakcija odvija, koja može biti privatna ili profesionalna

Transparentnost sustava UI je jedan je od važnih elemenata u izgradnji povjerenja među korisnicima. Sustavi moraju jasno komunicirati na koji način donose odluke te koje su granice njihovih mogućnosti, što smanjuje osjećaj nesigurnosti i povećava povjerenje korisnika. Identificirani su slučajevi s problemima povjerenja u sustav koji ukazuju na potrebu za većom transparentnošću i odgovornošću u dizajnu sustava UI [61].

Isto tako, Transparentnost se također odnosi i na svijest ljudi o interakciji sa sustavima UI. Kada su ljudi svjesni da komuniciraju s UI, prilagođavaju svoje ponašanje. Na primjer, ako znaju da razgovaraju s chatbotom, koriste jednostavnije rečenice kako bi postigli željene rezultate. Međutim, premda je važno je da sustavi mogu objasniti kako donose odluke, objašnjivost je ključna u nekim slučajevima, ali ne u svim. Naime, iako postoji potreba za transparentnošću, često ju je teško postići zbog složenosti sustava UI ili ograničenih mogućnosti razumijevanja metoda i algoritama. Potpuno razumijevanje sustava UI nije uvijek potrebno, kao što koristimo telefone i televizore bez poznavanja svih tehničkih detalja. Stoga je, umjesto same objašnjivosti, važan faktor upravljanje očekivanjima, odnosno prikazivanje mogućnosti i ograničenja sustava UI na dovoljno transparentan način. Informacije o tome što sustav postiže, kako i s kojom kvalitetom rezultata povećavaju razumijevanje ljudi i određuju uspjeh interakcija između ljudi i sustava UI [60].

Nadalje, kako ne bi diskriminirali, štetili ili obmanjivali korisnike te kako bi poštivali njihovu privatnost, vrijednosti i prava, potrebno je osigurati da sustavi UI budu nepristrani i odgovorni. Sustavi trebaju omogućiti korisnicima preuzimanje kontrole kad je to potrebno, dok UI pruža podršku kroz autonomne akcije kada je to prikladno. Obzirom da se tehnologije umjetne inteligencije sve više razvijaju, tako se i metodologije istraživanja HCI područja moraju prilagoditi, što uključuje prilagodbu postojećih tehnika evaluacije i razvoj novih metoda koje mogu učinkovito procijeniti interakciju sa sustavima UI [61], [62].

Da bi tehnologija bila uspješno prihvaćena od strane korisnika, ona mora ispuniti njihova očekivanja i potaknuti ih da je nastave koristiti. Ovo je posebno izazovno kod umjetne inteligencije zbog velike medijske pažnje koju privlači. Ljudi često zanemaruju činjenicu da i sami često griješe u obavljanju zadataka. Skepticizam prema sustavima UI prisutan je jer mnogi vide UI kao konkurenciju, osobito na profesionalnom planu, što može utjecati na njihovo samopouzdanje [60].

U tom kontekstu, važno je razlikovati prihvaćanje nakon početnog korištenja od namjere nastavka korištenja. Model prihvaćanja tehnologije (engl. *Technology Acceptance Model, TAM*) identificira dva ključna razloga prihvaćanja tehnologije od strane korisnika – percipiranu korisnost (engl. *perceived usefulness*) i percipiranu jednostavnost korištenja (engl. *perceived ease of use*). Iako je percipirana jednostavnost važna za početno korištenje, ona nije ključna za nastavak. Zadovoljstvo korisnika tehnologijom određuje njihovu namjeru da je nastave koristiti, potvrđujući očekivanja korisnika. Korisnici će tolerirati određene poteškoće u korištenju sustava ako pruža kritičnu funkcionalnost koja povećava njegovu korisnost. Međutim, nikakva količina jednostavnosti korištenja ne može nadoknaditi sustav koji ne pruža nikakve korisne prednosti [63].

Na prihvaćanje sustava UI mogu utjecati i izgled sustava te ponašanje koje nalikuje na ljudsko. Međutim, antropomorfizam, odnosno dodavanje ljudskih osobina sustavima UI, može imati dvojak učinak na korisničko iskustvo, ovisno o kontekstu interakcije i karakteristikama krajnjeg korisnika. S jedne strane, ljudski izgled sustava povećava angažiranost korisnika i pozitivno utječe na korisnikovu percepciju korisnosti i povjerenje u sustav UI. S druge strane, ljudske karakteristike sustava UI mogu negativno utjecati na interakciju kada očekivanja korisnika nisu ispunjena. Dakle, ovisno o kontekstu, ljudske osobine sustava UI mogu biti više ili manje potrebne. Ako sustavi UI djeluju kao pružatelji informacija, manje je potrebno da izgledaju ljudski nego kada djeluju kao prijatelji s društvenim i emocionalnim karakteristikama, gdje je veći angažman u interakciji važan. Antropomorfizam izravno utječe na očekivanja ljudi o sposobnostima sustava UI i njihovoj interakciji s njima. Kada sustav UI ima ljudske karakteristike, ljudi imaju ljudska očekivanja, a sustavi poput Siri su još uvijek daleko od ljudskih sposobnosti i imaju ograničene mogućnosti. Ako ta očekivanja postanu prevelika, postoji rizik od razočaranja, frustracije i konačno odbacivanja sustava UI. Štoviše, stvaranje ljudskih i prirodnih efekata također može povećati strah ljudi da postanu suvišni [60].

Postoje različita područja primjene interakcije sa sustavima umjetne inteligencije. Jedno od njih je kolaborativno donošenje odluka, gdje UI pruža uvid i analizira podatke kako bi podržala ljudske odluke u sektorima poput zdravstva, financija i poslovanja. Sljedeće područje su virtualni asistenti kao npr. Siri, Alexa ili Google Assistant, primjer interakcije čovjeka i UI koja omogućuje korisnicima obavljanje zadataka i pristup informacijama putem glasovnih naredbi. Zatim postoje personalizirani sustavi preporuka koji koriste algoritme UI za analizu preferencija i ponašanja korisnika, što je često prisutno u streaming servisima, online kupovini i društvenim mrežama. Obrada prirodnog jezika omogućuje sustavima UI razumijevanje i interpretaciju ljudskog jezika, što je ključno za chatbote, glasovne asistente i alate za prevođenje jezika. U autonomnim vozilima, UI analizira podatke senzora i donosi odluke u stvarnom vremenu, pomažući u navigaciji i izbjegavanju sudara. U zdravstvu, UI pomaže u postavljanju dijagnoze, planiranju liječenja i praćenju pacijenata, poboljšavajući točnost i dostupnost medicinskih usluga. U obrazovanju, UI nudi personalizirano podučavanje, podržava učenike i pruža vrijedne alate edukatorima. Ove primjene predstavljaju samo dio potencijala interakcije čovjeka i UI, a daljnji razvoj tehnologije donijet će još inovativnijih rješenja u različitim područjima [62].

## Taksonomija upotrebe umjetne inteligencije

Taksonomija upotrebe UI može poboljšati razumijevanje strukture interakcija između čovjeka i sustava UI te pomoći u evaluaciji i mjerenju tih interakcija za postizanje optimalnih rezultata. Razvoj taksonomije, neovisne o tehnikama i domenama UI, zahtijeva duboko razumijevanje uloge sustava UI u postizanju željenih ishoda jer je fokus na ishodu ili rezultatima interakcije korisnika s umjetnom inteligencijom sustava, umjesto da se fokusira na način na koji se izvršava interakcija korisnika. Taksonomija koju prikazuje Tablica 4.1 se sastoji od detaljnog seta aktivnosti i njihovih definicija, koje opisuju funkcije sustava UI u doprinosu ciljevima korisnika [64].

Tablica 4.1 Taksonomija upotrebe umjetne inteligencije [64]

|  |  |  |
| --- | --- | --- |
| **Aktivnost** | **Opis** | **Primjeri korištenja** |
| Stvaranje sadržaja | generiranje novih sadržaja poput videa, teksta, programskog koda, sintetičkih podataka | stvaranje prijevoda; stvaranje slike na temelju teksta |
| Sinteza sadržaja | kombiniranje i/ili sažimanje dijelova, elemenata ili koncepata u koherentnu cjelinu | konvertiranje neuređenih bilješki liječnika; pisanje sažetka knjige |
| Donošenje odluka | odabir između mogućih alternativa kako bi se došlo do rješenja | kupnja/prodaja, financijske odluka |
| Otkrivanje | identificiranje, pažljivim pretraživanjem, pregledavanjem ili istraživanjem, postojanja ili prisutnosti nečega | otkrivanje prijetnji kibernetičkoj sigurnosti |
| Digitalna asistencija | osobni agent za razumijevanje i odgovaranje na naredbe i pitanja te izvršavanje zadataka po nalogu | podsjetnici iz pametnih asistenta (npr. Siri, Amazon Echo, Google Assistant, Bixby) |
| Otkrivanje | pronalazak, prepoznavanje nečega po prvi put | otkrivanje i proizvodnja lijekova |
| Analiza slike | prepoznavanje atributa u digitalnim slikama za izvlačenje smislenih informacija | medicinska dijagnostika |
| Pronalaženje informacija | pronalaženje informacija o specifičnim temama od interesa | ubrzanje pretrage stabilnih proteina korištenih u razvoju lijekova i proizvodnji hrane |
| Praćenje | promatranje, provjeravanje i nadziranje procesa, kvalitete ili stanja tijekom vremena | praćenje šumskih požara |
| Poboljšanje performansi | poboljšanje kvalitete i učinkovitosti rezultata | analiza grafova, povećanje učinkovitosti i skalabilnosti |
| Personalizacija | dizajniranje i prilagođavanje nečega karakteristikama, preferencijama ili ponašanju pojedinca | personalizacija i analitika sadržaja prodaje |
| Predviđanje | prognoza vjerojatnosti budućeg ishoda | prognoza prodaje; vremenska prognoza |
| Automatizacija procesa | obavljanje ponavljajućih zadataka, smanjivanje grešaka i gubitka podataka te povećanje učinkovitosti procesa | automatizacija administrativnih zadataka |
| Preporuke | predlaganje opcija za pomoć u donošenju odluka | prijedlozi odgovora službe za korisnike; preporuke za kupnju; preporuke sadržaja |
| Robotička automatizacija | korištenje fizičkih strojeva za automatizaciju, poboljšanje i/ili optimizaciju različitih zadataka | inteligentni roboti u kirurgiji |
| Automatizacija vozila | Automatizacija fizičkog prijevoza robe, instrumentacije i/ili ljudi | autonomni automobili, kamioni, vlakovi; dronovi, svemirske letjelice; avioni |

## Umjetna inteligencija usmjerena na čovjeka

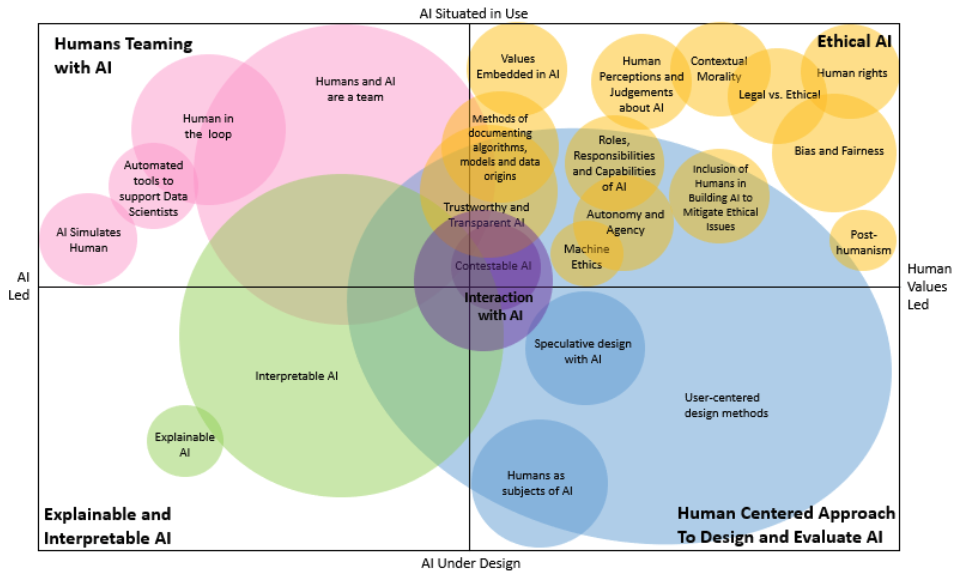
Umjetna inteligencija usmjerena na čovjeka (engl. *Human-Centered AI, HCAI*) predstavlja novu disciplinu usmjerenu na stvaranje sustava UI koji pojačavaju i unapređuju ljudske sposobnosti, umjesto da ih zamjenjuju. Cilj je očuvati kontrolu nad umjetnom inteligencijom te osigurati da UI zadovoljava naše potrebe, radi transparentno, pruža pouzdane rezultate i poštuje privatnost. Također se istražuju i oblikuju novi oblici interakcije između ljudi i sustava UI koji poboljšavaju i proširuju ljudske sposobnosti u korištenju tih sustava. Prihvaćanjem temeljne vrijednosti da su čovjek i sustav UI zajedno bolji od bilo kojeg pojedinačno, teži se razvoju nove korisničke interakcije koja potiče suradnju između ljudi i umjetne inteligencije. Također, mogu se stvoriti okviri za oblikovanje ili evaluaciju modela interakcije sa sustavima UI te se može provoditi teorijski rad koji razvija i proširuje teorije suradnje ili zajedničkog stvaranja ljudi i sustava UI. Primjer ovakve suradnje je podatkovna znanost, gdje UI pomaže znanstvenicima u analizi velikih i raznolikih skupova podataka i omogućuje stjecanje novih uvida i rješavanje složenih problema [65].

HCAI obuhvaća sve aspekte kako sustavi umjetne inteligencije usmjereni na čovjeka mogu pružiti pozitivne i korisne ishode svojim korisnicima, onima na koje utječe njihovo djelovanje te društvu općenito. Potrebno je razumjeti kako ljudi koriste i vjeruju sustavima UI. Također je potrebno objasniti kako sustavi UI funkcioniraju. HCAI se uz to bavi i procjenom negativnih posljedica ili potencijalne zloupotrebe sustava UI i analizom percepcije ljudi o sustavima UI. HCAI je u skladu sa konceptom etičke umjetne inteligencije i naglašava važnost poštivanja ljudskih prava, pravednosti i raznolikosti. Naglašava se dodjeljivanje odgovornosti za radnje i odluke koje sustavi UI donose, što osigurava da su pojedinci ili organizacije odgovorni za rezultate sustava. Ta odgovornost obuhvaća cijeli životni ciklus, od oblikovanja i uvježbavanja modela UI do implementacije i nadzora sustava. Ako je sudionicima dodijeljena odgovornost, potiče ih se na davanje prioriteta pravednosti, jednakosti i etičkoj upotrebi. Pristup koji se temelji na odgovornosti ključan je za izgradnju čvrstih etičkih temelja u sustavima UI [65], [66].

Razumijevanje potreba, izazova i konteksta korisnika je od presudne važnosti. Potrebno je stvoriti sustave UI koji zaista rješavaju njihove probleme i poboljšavaju njihov život. Uključivanje korisnika u proces razvoja ključno je za stvaranje sustava koji su istinski korisni i prilagođeni korisnicima. Ovakav pristup osigurava da su rješenja usklađena sa stvarnim potrebama i željama. Osim toga, sustavi UI trebaju biti dostupni i upotrebljivi za što širi krug ljudi, bez obzira na njihove mogućnosti ili podrijetlo. Potrebno je osigurati da prednosti koje pruža UI budu dostupne svima. HCAI zahtijeva iterativni proces koji uključuje stalno testiranje, povratne informacije i unaprjeđivanje. Ovakav pristup osigurava da se sustavi UI razvijaju u skladu s promjenjivim potrebama korisnika i tehnološkim napretkom. Iako UI može automatizirati mnoge zadatke, važno je održati ravnotežu u kojoj ljudi ostaju u kontroli, posebno u kritičnim situacijama donošenja odluka. Ova ravnoteža osigurava da UI nadopunjuje, a ne zamjenjuje ljudske sposobnosti [66].

Sve veći broj ljudi sve češće stupa u interakciju sa sustavima UI ili je pod njihovim utjecajem u raznim područjima. Razumijevanje percepcija UI među općom populacijom može pomoći u usmjeravanju istraživanja o interakciji između ljudi i sustava UI te podržati razvoj sustava usmjerenih na čovjeka. Povjerenje se pokazalo kao važan faktor u interakciji čovjeka i sustava UI. Mjerenje povjerenja može pomoći u pojašnjavanju učinaka varijabli vezanih uz korisnike, sustave i ponašanje korisnika prema tehnologiji. Nedavna nastojanja u smjeru pouzdane i odgovorne UI naglašavaju važnost pristupa koji uzimaju u obzir subjektivna iskustva onih koji koriste ili su pod utjecajem sustava UI [67].

Postoje četiri glavna područja istraživanja pod okriljem HCAI kao što prikazuje Slika 4.1. To su: Objašnjiva i interpretabilna UI (engl. *Explainable and Interpretable AI*), Pristupi oblikovanja i evaluacije UI usmjereni na čovjeka (engl. *Human Centered Approach to Design and Evaluate AI*), Udruživanje ljudi i UI (engl. *Humans Teaming with AI*) te Etička UI (engl. *Ethical AI*). Peto, novije područje koje se nalazi u središtu mape je Interakcija s umjetnom inteligencijom [68].



Slika 4.1 Područja istraživanja u domeni HCAI, preuzeto sa [68]

Objašnjiva i interpretabilna UI obuhvaća alate, metode i okvire koji pomažu ljudima u razumijevanju odluka i predviđanja sustava UI. Ova istraživačka područja razvijena su kao odgovor na koncept "crne kutije" modela UI, gdje često nije jasno kako se donose odluke.

Pristup usmjeren na ljude u dizajnu i evaluaciji UI koristi metode i alate HCI područja. Ovi pristupi uključuju specifične metode oblikovanja, studije i evaluacije s krajnjim korisnicima.

Udruživanje ljudi i UI omogućuje postizanje boljih rezultata nego što bi to bilo moguće samostalnim radom. Ovaj koncept se proučava kroz prizmu suradnje, često u kontekstu *human in-the-loop* koncepta gdje su ljudi uključeni u stvaranje i korištenje modela strojnog učenja, što može uključivati označavanje podataka, treniranje modela, ispravljanje pogrešnih klasifikacija i evaluaciju. Ovakva suradnja može značajno poboljšati performanse sustava UI, jer doprinosi u kontinuiranom razvoju modela, što je nešto drukčija uloga čovjeka od one koju zagovara etička UI gdje je fokus više na širim društvenim i etičkim pitanjima [68].

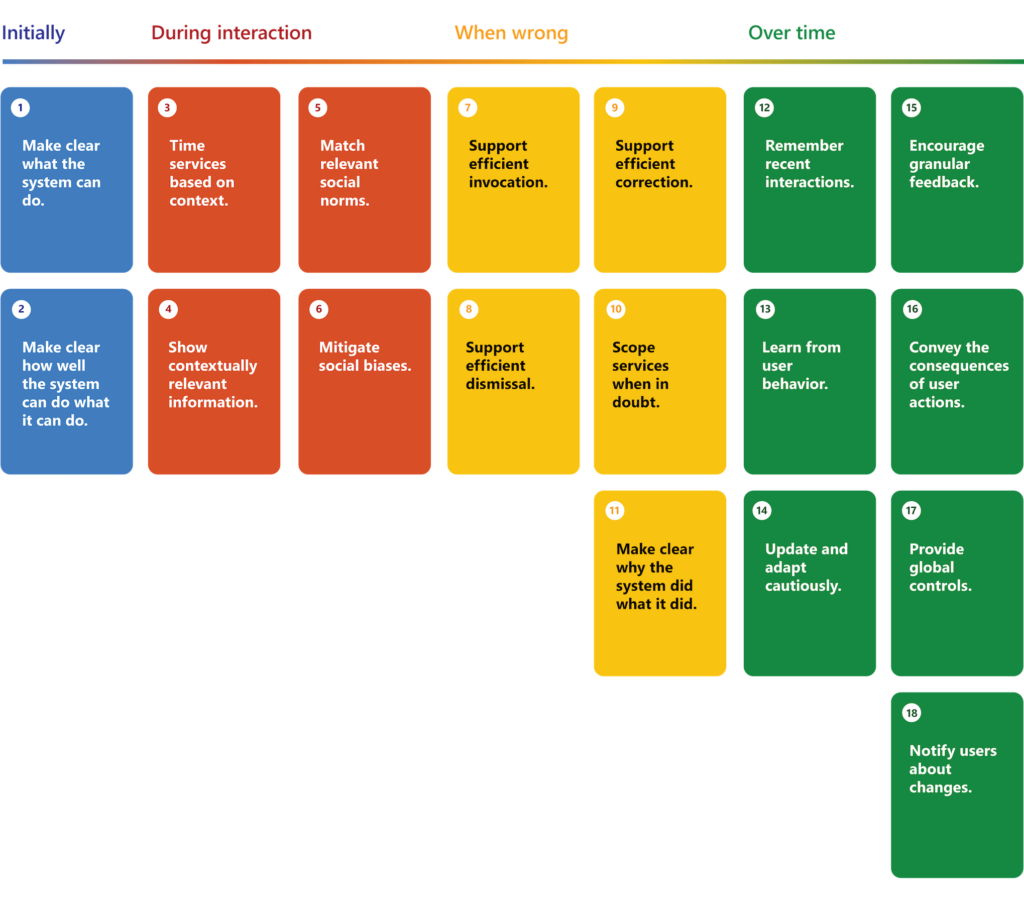
Interakcija s UI obuhvaća istraživanje kako ljudi koriste, preispituju i prilagođavaju sustave UI u svakodnevnom kontekstu. Ovo područje istražuje načine na koje sustavi UI mogu biti dizajnirani da bolje odgovaraju korisničkim potrebama. Istraživanja se usredotočuju na vraćanje korisničke autonomije, otkrivanje korisničkih potreba i poboljšanje dizajna interakcije. Ključni ciljevi uključuju dizajn koji omogućuje bolje razumijevanje i upotrebu UI sustava te isticanje etičkih pitanja i odgovornosti. To zahtijeva multidisciplinarne timove koji kombiniraju tehničke, dizajnerske i korisnički orijentirane pristupe. Kao odgovor na to Međunarodna tijela kao što je Europska unija počela su se baviti implikacijama umjetne inteligencije u društvu, zalažući se za stvaranje umjetne inteligencije usmjerene na čovjeka. Nadalje sveučilišta Stanford, UC Berkeley i MIT uspostavili su istraživačke institute HCAI, zagovarajući umjetnu inteligenciju koja je humanistička i etička [68].

Iako umjetna inteligencija može preuzeti donošenje odluka u rutinskim i strukturiranim zadacima, nerealno je očekivati da će UI preuzeti vodstvo u svim profesionalnim kontekstima. Kompleksnost problema, ljudski faktori, odgovornost za odluke, kontekstualna nejasnoća i neizvjesnost u odlučivanju preveliki su izazovi za trenutačno stanje UI. Unatoč rastućem interesu za interakciju između ljudi i umjetne inteligencije, postoji potreba za dubljom teorijom i istraživanjem koja bi pružila bogatiju konceptualizaciju ljudskog iskustva korištenja sustava UI u različitim kontekstima, kako iz tehničke tako i iz ne-tehničke perspektive. Naime, razvoj sustava UI usmjeren je na stvaranje učinkovitijih i jednostavnijih korisničkih sučelja. Međutim, postoje različita stajališta o pouzdanosti sustava UI te je učestalo mišljenje da je njihovo sučelje samo kozmetičko, što dovodi do neslaganja oko smjernica i principa za korisnička sučelja sustava UI. Osim toga, potrebno je istražiti kako već poznate zamke ljudske komunikacije mogu utjecati na komunikaciju između ljudi i sustava UI te kako izbjeći potencijalne nesporazume. Jedan od problema je i to što instrumente objašnjive UI često razvijaju računalni znanstvenici za računalne znanstvenike umjesto da se više usredotoče na krajnje korisnike koji zahtijevaju pristupačnije oblike objašnjenja [69].

Na posljetku kako sustavi UI postaju sve učinkovitiji i autonomniji, oblici interakcije s tim sustavima se moraju transformirati. Da bi se učinkovito podržala suradnja između ljudi i sustava UI, važno je omogućiti sustavu i korisniku da pregovaraju o zajedničkom cilju. Često je pri izradi sustava UI fokus na optimizaciji performansi ili točnosti, no ti parametri ne obuhvaćaju druge ciljeve poput personalizacije i pravednosti, koje ljudi uzimaju u obzir pri donošenju odluka. Predviđa se da će ljudi preuzeti ulogu specifikacije, postavljanja ciljeva, usmjeravanja, visokog stupnja kreativnosti i upravljanja, dok će UI nadopunjavati ljudske sposobnosti kroz inspiraciju, kreativnost, rad na detaljima i sposobnost dizajniranja u velikom obimu. Dakle, unatoč napretku, očekivano je da će sustavi UI uvijek zahtijevati ljudsku intervenciju za upravljanje društvenim aspektima i donošenje odluka. Kao što je već spomenuto, cilj je poboljšati ljudske sposobnosti, a ne ih zamijeniti [65], [70].

## Smjernice za oblikovanje interakcije sa sustavima umjetne inteligencije

Napredak u području UI pokrenuo je njenu primjenu u korisničkim aplikacijama i uslugama. Uz nove prilike nastali su i izazovi oblikovanja sustava UI usmjerenih na korisnika koji su potakli istraživače iz tvrtke Microsoft na sintetiziranje smjernica za oblikovanje interakcije između ljudi i sustava UI. Dokument sa smjernicama nastao je na temelju brojnih preporuka za dizajn interakcije s sustavima UI koje su znanstvenici i inženjeri predlagali u periodu od prethodnih 20 godina. Preko 150 takvih preporuka grupirano je i sažeto u jedinstven set smjernica. Nakon evaluacije sa stručnjacima za HCI, smjernice su postale lako razumljive i primjenjive na širok raspon proizvoda. Predloženih 18 smjernica za interakciju između ljudi i sustava UI grupirano je u četiri kategorije, kao što prikazuje Slika 4.2, koje definiraju ponašanje sustava UI pri početnoj interakciji, tijekom interakcije, u slučaju kada sustav pogriješi, i tijekom vremena. Smjernice nisu zamišljene kao jednostavna kontrolna lista, već kao potpora pri oblikovanju sustava i poticaj za daljnja istraživanja [71].



Slika 4.2 Smjernice za oblikovanje interakcije sa sustavima UI, preuzeto sa [72]

Svaka smjernica je napisana kao kratko pravilo uz koje ide opis od jedne rečenice koji pojašnjava potencijalne dvosmislenosti. Pravilo ne sadrži veznike kako bi se za konkretni primjer sučelja moglo jasno potvrditi je li pravilo zadovoljeno ili ne. Za svako pravilo dan je i primjer upotrebe, a u dodatku rada navedena su tri primjera u kojima se pravilo poštuje i tri primjera u kojima je ono prekršeno [73].

1. **Razjasniti što sustav može učiniti.** Pomoći korisniku razumjeti mogućnosti sustava.
2. **Razjasniti koliko dobro sustav može učiniti ono što može učiniti.** Pomoći korisniku da razumije koliko često sustav UI može pogriješiti.
3. **Vrijeme usluge prema kontekstu.** Uskladiti vrijeme kada treba djelovati ili prekinuti prema trenutnom zadatku i okruženju korisnika.
4. **Prikazati kontekstualno relevantne informacije.** Prikazati informacije relevantne trenutnom zadatku korisnika i okruženju.
5. **Uskladiti se s relevantnim društvenim normama.** Osigurati korištenje sustava na način koji korisnici očekuju, uzimajući u obzir njihov društveni i kulturni kontekst.
6. **Ublažiti društvene predrasude.** Osigurati da jezik i ponašanje sustava UI ne pojačavaju nepoželjne i nepravedne stereotipove i predrasude.
7. **Omogućiti efikasno pozivanje.** Olakšati pozivanje ili zahtijevanje usluga sustava UI kada je potrebno.
8. **Omogućiti efikasno odbacivanje.** Olakšati odbacivanje ili ignoriranje nepoželjnih usluga sustava UI.
9. **Omogućiti efikasnu korekciju.** Olakšati uređivanje, usavršavanje ili oporavak kada sustav UI pogriješi.
10. **Ograničiti usluge u slučaju sumnje.** Baviti se višeznačnošću ili ograničiti usluge sustava UI kada ste nesigurni u ciljeve korisnika.
11. **Razjasniti zašto je sustav radio na određeni način.** Omogućiti korisniku da pristupi objašnjenju zašto se sustav UI ponašao na određeni način.
12. **Zapamtiti nedavne interakcije.** Održavati kratkoročno pamćenje i omogućiti korisniku da se poziva na prethodne interakcije.
13. **Učiti iz ponašanja korisnika.** Personalizirati korisnikovo iskustvo učenjem iz njihovih radnji tijekom vremena.
14. **Osigurati oprezno ažuriranje sustava. Ograničiti uznemiravajuće promjene pri ažuriranju sustava UI.**
15. **Poticati povratnu informaciju.** Omogućiti korisniku da pruži povratne informacije koje indiciraju njegove preferencije tijekom regularne interakcije sa sustavom UI.
16. **Priopćiti posljedice korisničkih radnji.** Odmah priopćite kako će korisničke radnje utjecati na buduće ponašanje sustava UI.
17. **Omogućiti globalne kontrole.** Omogućiti korisniku prilagoditi što sustav UI prati i kako se ponaša.
18. **Obavijestiti korisnike o promjenama.** Obavijestite korisnika kada sustav UI doda ili ažurira svoje sposobnosti.

Relevantnost smjernica potvrđena je kroz korisničku studiju s 49 stručnjaka s potrebnim iskustvom koji su testirali smjernice na 20 sustava temeljenih na UI. Identificirano je više potencijalnih problema, a jedan od njih je činjenica da sustavi temeljeni na UI mogu biti nepredvidivi i nekonzistentni, što može dovesti do zabune korisnika i potencijalne štete. Npr. sustavi koji se temelje na UI mogu različito reagirati ovisno o rasvjeti ili prisutnosti buke u okolini. Također sustavi mogu različito reagirati na isti tekstualni upit tijekom vremena zbog ažuriranja jezičnog modela koji se koristi ili se ponašati različito od jednog do drugog korisnika zbog personalizacije. Problem je i to što tradicionalne smjernice za upotrebljivost tehnologije nisu uvijek primjenjive. Kao nedostatak prethodnih istraživanja navodi se da prijedlozi za oblikovanje sustava UI često nisu dovoljno eksplicitni i da se u mnogim slučajevima problemi upotrebljivosti sustava UI i moguća rješenja navode tek u raspravi ili najavi budućih radova [71].

## Agentski sustavi temeljeni na umjetnoj inteligenciji

Agentom smatramo entitet koji može djelovati u skladu s ljudskim namjerama. Osnovne sposobnosti koje agent mora imati su slanje poruka, primanje poruka, odgovaranje drugim agentima nakon poduzetih radnji i komunikacija s njima. Agenti mogu koristiti različite tipove sustava za podršku u obavljanju radnji i generiranju odgovora. Npr. neki agenti koriste modele UI za generiranje odgovora, dok drugi koriste funkcije sustava temeljenih na alatima, a neki koriste ljudski unos za odgovaranje drugim agentima. Kao što prikazuje Slika 4.3, moguće je imati agente koji kombiniraju različite tipove sustava ili kompleksnije agente koji imaju unutarnje razgovore među više agenata, dok ih na površini drugi agenti doživljavaju kao jedinstvenu cjelinu za komunikaciju [74].

A diagram of a chat conversation

Description automatically generated

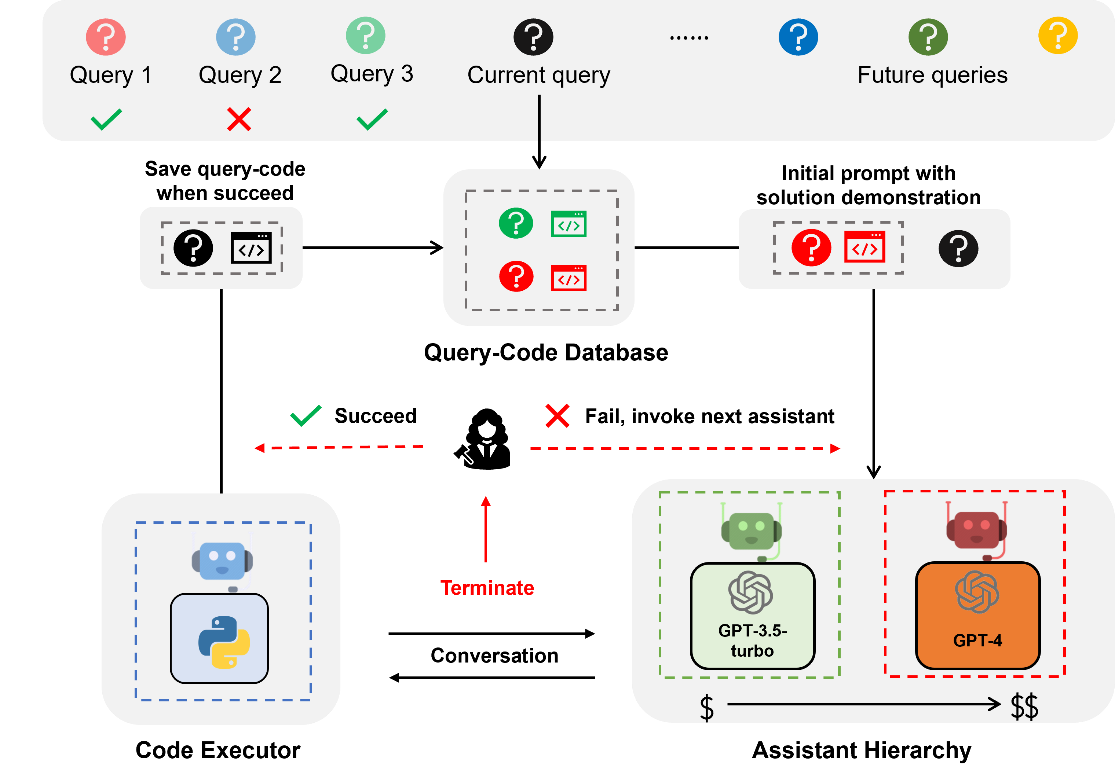
Slika 4.3 Višeagentska komunikacija, preuzeto sa [75]

Jedan od primjera agentskog sustava temeljenog na umjetnoj inteligenciji su LLM agenti, napredni sustavi UI koji koriste jezične modele za razumijevanje i generiranje ljudskog jezika na sofisticiran način i unutar određenog konteksta. To im daje mogućnost praćenja tijeka razgovora, prisjećanja prethodnih izjava te prilagodbe odgovora. LLM agent~~i~~ se sastoji od četiri osnovne komponente koje zajedno omogućuju obavljanje raznovrsnih zadataka i interakcija. Središnja komponenta interpretira ulazne podatke, primjenjuje zaključivanje te na temelju ciljeva agenta određuje najprikladniju akciju. Memorijska komponenta služi za pohranu prethodnih razgovora i kontekstualnih informacija za prilagođene i relevantne odgovore. Komponenta alata omogućuje izvođenje specifičnih zadataka poput izvršavanja programskog koda ili pretraživanja informacija, dok modul za planiranje omogućuje rješavanje kompleksnih problema, pomažući agentu da strategijski pristupi dugoročnim ciljevima i izazovima. Arhitektura LLM agenata sastoji se od samog LLM modela, integracijskog sloja za komunikaciju s drugim sustavima, slojeva za obradu ulaza i izlaza radi boljeg razumijevanja i odgovaranja te etičkih i sigurnosnih slojeva koji sprječavaju zloupotrebu i osiguravaju etički usklađene odgovore. Konačno, sučelje za korisnike omogućuje interakciju s ljudima putem tekstualnih, glasovnih ili čak robotskih sučelja, prilagođenih različitim vrstama interakcije [76].

Suprotno pojedinačnim LLM agentima, višeagentski sustavi temeljeni na LLM-ovima predstavljaju okvire u kojima više LLM agenata međusobno komunicira ili surađuje kako bi postigli kompleksne ciljeve. U ovakvom sustavu agenti nisu samo opremljeni sposobnošću učenja i prilagodbe, već i mogućnošću kontinuirane komunikacije s drugim agentima te pristupom raznovrsnim resursima bez ograničenja kontekstualnih prozora. Ovaj pristup proširuje mogućnosti pojedinačnih LLM agenata jer komunicirajući, surađujući i dijeleći informacije, višeagentski sustavi mogu učinkovitije rješavati probleme nego što bi to mogao pojedinačni agent [77].

Kada govorimo o višeagentskim sustavima potrebno je uzeti u obzir oblik sučelja, odnosno komunicira li korisnik sa sustavom putem jedne točke interakcije ili vidi više agenata s kojima mora pojedinačno komunicirati. Jedinstvena točka interakcije može pojednostaviti korisničko iskustvo mnogih aplikacija. Međutim, u nekim situacijama to nije najprikladnije rješenje. Na primjer, kada aplikacija uključuje raspravu više agenata o određenoj temi, korisnici trebaju vidjeti što svaki agent kaže. U takvim slučajevima korisnicima je korisno pratiti ponašanje više agenata. Druga važna dimenzija je arhitektura, koja se odnosi na to postoje li u pozadini sustava više agenata koji rade zajedno. Višeagentski sustav je lakše održavati, razumjeti i proširivati u usporedbi sa sustavom temeljenim na jednom agentu. Čak i za sučelja temeljena na jednom agentu, implementacija s više agenata može učiniti sustav modularnijim i olakšati programerima dodavanje ili uklanjanje funkcionalnih komponenti. Zapravo je višeagentska arhitektura dobar način za izgradnju sustava s jednim agentom. Počevši od jednostavnih agenata, moguće je njihovo učinkovito komponiranje i koordinacija kako bi se postigla viša razina inteligencije. Trenutno ne postoji dovoljno dobar agent koji može obaviti sve željene zadatke, a to može biti posljedica toga što još nije pronađen pravi način za komponiranje više agenata u moćnog pojedinačnog agenta [74].

Čak i kad je izvedba jednog agenta dovoljno dobra, postoji mogućnost smanjenja troškova upotrebom više agenata s jeftinijim modelom koja čak može i nadmašiti performanse jednog agenta sa skupljim modelom. Sustav EcoAssistant kojeg prikazuje Slika 4.4, dobar je primjer kombiniranja GPT-4 i GPT-3.5 agenata za smanjenje troškova uz poboljšanje performansi za određene zadatke, čak i u usporedbi s korištenjem samo jednog GPT-4 agenta [78].



Slika 4.4 Višeagentski sustav EcoAssistant, preuzeto sa [78]

Sustav se sastoji od LLM asistenata koji komuniciraju s automatskim izvršiteljem koda i mogu iterativno dorađivati kod ili proizvesti odgovore na temelju rezultata izvršenja koda. Postoji hijerarhija LLM asistenata koji pokušavaju odgovoriti na upit. Najprije idu oni sa slabijim, jeftinijim LLM-ovima prije nego što se pozovu jači, skuplji. Također se dohvaćaju rješenja iz prošlih uspješnih upita kao primjeri u kontekstu za nove upite [79].

Drugi primjer u kojem se korištenje višeagentskih sustava pokazalo učinkovitije je u području razvoja softvera. Sustav MAGIS koristi četiri agenta (upravitelj, čuvar repozitorija, programer i inženjer za osiguranje kvalitete) za zajedničko rješavanje problema na razini repozitorija koji se pojavljuju na GitHub-u i nadmašuje pojedinačno korištenje LLM-ova [80].

Još jedan važan aspekt koji se može unaprijediti korištenjem višeagenstkih sustava je memorija. Standardni asistenti temeljeni na velikim jezičnim modelima mogu zapamtiti trenutni razgovor s korisnikom te pokazati sposobnost učenja konteksta iz korisničkih smjernica tijekom razgovora. Međutim, memorija i naučeni kontekst asistenta se gube nakon završetka razgovora ili ako razgovor postane predugačak pa je korisnik prisiljen ponavljati instrukcije. Ovaj problem se može riješiti tako da se korisničke instrukcije spremaju izvan razgovora u dugoročnu memoriju implementiranu poput baze podataka. Umjesto kopiranja cijele memorije u kontekstni prozor, što bi zauzimalo dragocjeni prostor, pojedine instrukcije povlače se u kontekst po potrebi [81].

Na posljetku, korištenje višeagentskih sustava temeljenih na UI otvara i neka etička pitanja. Kao prvo, važno je osigurati zaštitu privatnosti i podataka korisnika tijekom interakcija između agenata. Zatim, korištenje jezičnih modela unutar sustava može rezultirati implicitnim pristranostima prisutnim u skupovima podataka na kojima su ti modeli trenirani. Nadalje, obzirom na složenost interakcija između više agenata, bitno je uspostaviti mehanizme jasne odgovornosti i transparentnosti. Korisnicima treba omogućiti razumijevanje i praćenje procesa donošenja odluka kako bi se osigurala odgovornost i rješavanje eventualnih problema [82].

## Bihevioristički pristup u analizi sustava umjetne inteligencije

Biheviorizam je jedna od temeljnih teorija učenja koja nudi zanimljive poglede i metodologije koje se mogu primijeniti u području umjetne inteligencije posebice u kontekstu interakcije s velikim jezičnim modelima. U tom smislu, ponašanja sustava UI, odnosno odgovori LLM-ova na upite, slični su ljudskim ponašanjima u klasičnoj biheviorističkoj teoriji. Koriste se podražaji (upiti) kako bi se izazvali željeni odgovori od sustava UI. U slučaju kad su ti odgovori usklađeni s očekivanim ishodom, sustav UI se potiče na ponavljanje takvog ponašanja u sličnim budućim situacijama [83].

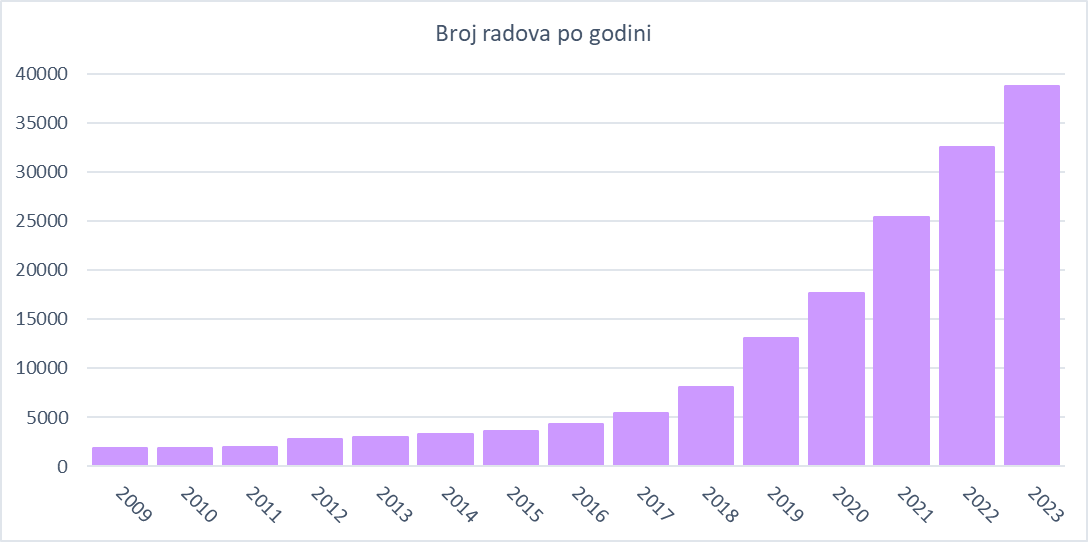
U biheviorizmu se naglasak stavlja na ponavljanje i potkrepljenje, a pri korištenju alata UI poput ChatGPT-a, mogu se prepoznati poželjna ponašanja i pružiti korektivne povratne informacije kada je to potrebno s ciljem stvaranja okruženja u kojem korisnici mogu prakticirati i usavršavati svoja ponašanja kroz kontinuiranu interakciju sa sustavom UI, dobivajući pravovremene, relevantne i konstruktivne povratne informacije [84].

Primjena biheviorizma ovisi o oblikovanju učinkovitih upita koji moraju biti jasni, sažeti i strukturirani na način koji vodi sustav UI prema željenom odgovoru. Potrebno se osvrnuti i na ograničenja biheviorizma, posebno na njegovu usmjerenost isključivo na vanjska ponašanja bez razmatranja unutarnjih misaonih procesa, što postaje sve važnije kako sustavi UI postaju napredniji i sposobniji oponašati složene ljudske kognitivne funkcije [83].

Bihevioristički pristup može biti koristan za razumijevanje i oblikovanje interakcija sa sustavima UI koji su postali vrlo kompleksni, s milijardama parametara unutar modela. Umjesto pokušaja razumijevanja tih složenih unutarnjih mehanizama, biheviorizam se fokusira na promatranje i manipulaciju vanjskih ponašanja sustava. Ovaj pristup omogućuje stvaranje učinkovitijih sustava kroz iterativno testiranje i prilagodbu interakcija temeljenih na opaženim odgovorima. U kontekstu velikih jezičnih modela, bihevioristička metodologija može pomoći u identificiranju obrazaca i optimizaciji odgovora sustava na temelju podražaja (upita) korisnika. Uzimajući u obzir nepredvidivost i dinamičnost ovih sustava, promatranje ponašanja i prilagodba strategija interakcije mogu pružiti praktična rješenja za poboljšanje učinkovitosti i točnosti sustava UI, dok se istovremeno omogućava sustavu da uči iz povratnih informacija i kontinuirano poboljšava svoje performanse.

# PREGLED PODRUČJA: INTERAKCIJA SA SUSTAVIMA UMJETNE INTELIGENCIJE TEMELJENIM NA VELIKIM JEZIČNIM MODELIMA

Umjetna inteligencija postala je jedno od najbrže rastućih područja znanosti čija popularnost raste eksponencijalno u posljednjih nekoliko godina. Kao što prikazuje Slika 5.1, broj znanstvenih članaka u bazi *Web of Science*, u čijem se naslovu ili sažetku spominje pojam umjetna inteligencija, povećao se u posljednjih deset godina s oko 3 tisuće na gotovo 40 tisuća. Ovaj rast naglašava sve veći interes i značaj umjetne inteligencije u suvremenoj znanosti i tehnologiji.

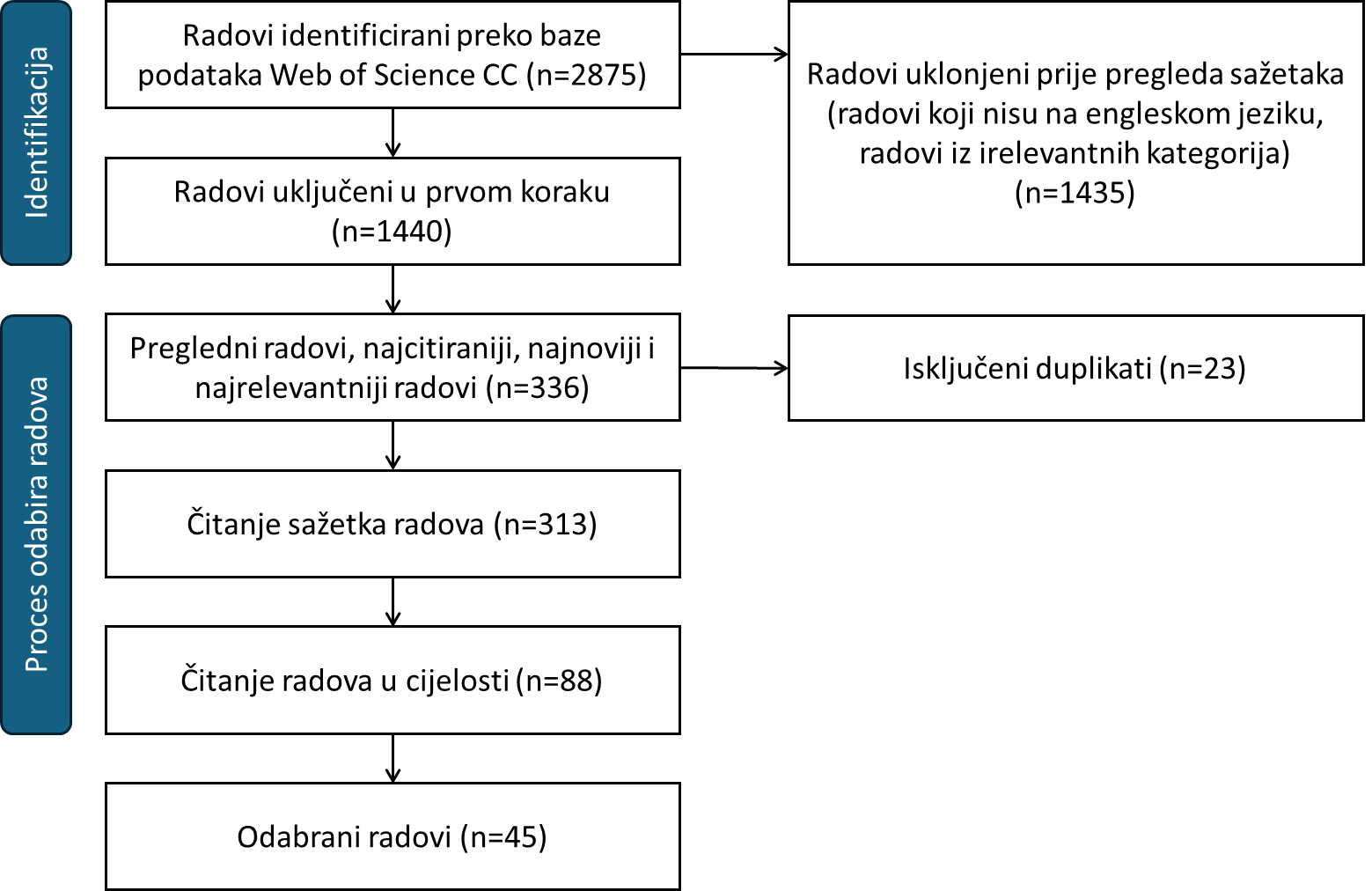


Slika 5.1 Broj publikacija u bazi Web of Science u čijem se naslovu ili sažetku spominje pojam umjetna inteligencija u razdoblju od 2009 do 2023 godine.

U prethodnom poglavlju opisana je interakcija sa sustavima umjetne inteligencije dok je u ovom poglavlju područje istraživanja suženo na samo one sustave koji se temelje na velikim jezičnim modelima. Pregled literature proveden je koristeći smjernice *Preferred Reporting Items for Systematic review and Meta-Analysis* (PRISMA) [85]. Sustavno je pretražena jedna od vodećih baza podataka *Web of Science Core Collection* (WoS CC) kako bi se pregledala literatura u nekoliko faza kao što prikazuje Slika 5.2.

Najprije je izvršena identifikacija relevantne literature, pri čemu su korištene određene ključne riječi obzirom na područje istraživanja (*LLM, large language model, GPT, ChatGPT, interact, prompt*). Uz riječi koje su se tražile unutar naslova i sažetaka radova kombinirani su Booleovi operatori „AND“ i „OR“ te specijalni znak „\*“ kako bi se što bolje obuhvatili relevantni rezultati, te je identificirano 2875 članaka iz spomenute baze podataka.

U sljedećem koraku identifikacije radovi su ograničeni na nekoliko relevantnih kategorija prema dosadašnjem smjeru istraživanja (*Computer Science, Education Educational Research, Psychology, Behavioral Sciences, Robotics*) što je smanjilo broj rezultata na 1440 radova.



Slika 5.2 PRISMA dijagram toka za odabir radova za pregled

Zatim su za daljnju analizu izabrani svi pregledni radovi te 100 najcitiranijih, 100 najnovijih i 100 najrelevantnijih radova po WoS kriteriju, što je dovelo do ukupne brojke od 313 radova na engleskom jeziku, bez duplikata.

Konačno, pregledom sažetaka izbačeni su radovi koji nisu povezani s temom istraživanja te je izdvojeno 88 radova od kojih je nakon čitanja cijelog teksta odabrano njih 45 koji su opisani dalje u tekstu. Radovi su raspodijeljeni u nekoliko WoS kategorija kao što prikazuje Slika 5.3.



Slika 5.3 Analiza rezultata pretrage, prikaz po WoS kategorijama

Nadalje, Tablica 5.1 sadrži osnovne podatke o 45 odabranih radova kao što su naslov rada, naziv časopisa ili konferencije u kojoj je rad objavljen, broj citata u WoS bazi te godina objave rada, dok su u tekstu nakon tablice navedeni i analizirani zaključci i rezultati odabranih radova.

Tablica 5.1 Pregled odabranih radova

|  |  |  |  |
| --- | --- | --- | --- |
| **Naslov rada** | **Naziv časopisa/konferencije** | **Broj citata** | **Godina objave** |
| Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis | CONNECTION SCIENCE | 0 | 2024 |
| LLM4VV: Developing LLM-driven testsuite for compiler validation | FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE | 0 | 2024 |
| Enhancing the LLM-Based Robot Manipulation Through Human-Robot Collaboration | IEEE ROBOTICS AND AUTOMATION LETTERS | 0 | 2024 |
| Towards Training-Free Open-World Segmentation via Image Prompt Foundation Models | INTERNATIONAL JOURNAL OF COMPUTER VISION | 0 | 2024 |
| When large language models meet personalization: perspectives of challenges and opportunities | WORLD WIDE WEB-INTERNET AND WEB INFORMATION SYSTEMS | 0 | 2024 |
| Explainable Knowledge reasoning via thought chains for knowledge-based visual question answering | INFORMATION PROCESSING & MANAGEMENT | 0 | 2024 |
| Examining How the Large Language Models Impact the Conceptual Design with Human Designers: A Comparative Case Study | INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION | 0 | 2024 |
| 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models | LEARNING MEDIA AND TECHNOLOGY | 0 | 2024 |
| RAH! RecSys-Assistant-Human: A Human-Centered Recommendation Framework With LLM Agents | IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS | 0 | 2024 |
| A Large Language Model Approach to Educational Survey Feedback Analysis | INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE IN EDUCATION | 0 | 2024 |
| ChatGPT: perspectives from human-computer interaction and psychology | FRONTIERS IN ARTIFICIAL INTELLIGENCE | 0 | 2024 |
| The impact of large language models on higher education: exploring the connection between AI and Education 4.0 | FRONTIERS IN EDUCATION | 0 | 2024 |
| 3A-COT: an attend-arrange-abstract chain-of-thought for multi-document summarization | INTERNATIONAL JOURNAL OF MACHINE LEARNING AND CYBERNETICS | 0 | 2024 |
| Human-computer pragmatics trialled: some (im)polite interactions with ChatGPT 4.0 and the ensuing implications | INTERACTIVE LEARNING ENVIRONMENTS | 0 | 2024 |
| Prompt Sapper: A LLM-Empowered Production Tool for Building AI Chains | ACM TRANSACTIONS ON SOFTWARE ENGINEERING AND METHODOLOGY | 1 | 2024 |
| Improving Training Dataset Balance with ChatGPT Prompt Engineering | ELECTRONICS | 0 | 2024 |
| Using ChatGPT in Software Requirements Engineering: A Comprehensive Review | FUTURE INTERNET | 0 | 2024 |
| A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions | COMPUTER SCIENCE REVIEW | 0 | 2024 |
| From Traditional Recommender Systems to GPT-Based Chatbots: A Survey of Recent Developments and Future Directions | BIG DATA AND COGNITIVE COMPUTING | 0 | 2024 |
| Unraveling the mysteries of AI chatbots | ARTIFICIAL INTELLIGENCE REVIEW | 0 | 2024 |
| Dawn of the dialogue: AI's leap from lab to living room | FRONTIERS IN ARTIFICIAL INTELLIGENCE | 0 | 2024 |
| The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach | FRONTIERS IN EDUCATION | 5 | 2024 |
| Diversity Measures: Domain-Independent Proxies for Failure in Language Model Queries | 18TH IEEE INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING, ICSC 2024 | 0 | 2024 |
| ChatGPT for Robotics: Design Principles and Model Abilities | IEEE ACCESS | 21 | 2024 |
| Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions | EDUCATIONAL TECHNOLOGY & SOCIETY | 29 | 2023 |
| Chat With ChatGPT on Interactive Engines for Intelligent Driving | IEEE TRANSACTIONS ON INTELLIGENT VEHICLES | 23 | 2023 |
| Unleashing the Potential of Generative AI, Conversational Agents and Chatbots in Educational Praxis: A Systematic Review and Bibliometric Analysis of GenAI in Education | OPEN PRAXIS | 4 | 2023 |
| Prompting Large Language Models With the Socratic Method | 2023 IEEE 13TH ANNUAL COMPUTING AND COMMUNICATION WORKSHOP AND CONFERENCE, CCWC | 4 | 2023 |
| Generating Requirements Elicitation Interview Scripts with Large Language Models | 2023 IEEE 31ST INTERNATIONAL REQUIREMENTS ENGINEERING CONFERENCE WORKSHOPS, REW | 2 | 2023 |
| Fairness-guided Few-shot Prompting for Large Language Models | ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 36 (NEURIPS 2023) | 0 | 2023 |
| A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model | JOURNAL OF WEB ENGINEERING | 0 | 2023 |
| Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems | HHAI 2023: AUGMENTING HUMAN INTELLECT | 1 | 2023 |
| Flexible and Secure Code Deployment in Federated Learning using Large Language Models: Prompt Engineering to Enhance Malicious Code Detection | 2023 IEEE INTERNATIONAL CONFERENCE ON CLOUD COMPUTING TECHNOLOGY AND SCIENCE, CLOUDCOM 2023 | 0 | 2023 |
| Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering | 2023 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR) | 22 | 2023 |
| Enabling Conversational Interaction with Mobile UI using Large Language Models | PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI 2023 | 15 | 2023 |
| Improved Trust in Human-Robot Collaboration With ChatGPT | IEEE ACCESS | 21 | 2023 |
| Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts | PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI 2023) | 47 | 2023 |
| Chat with the Environment: Interactive Multimodal Perception Using Large Language Models | 2023 IEEE/RSJ INTERNATIONAL CONFERENCE ON INTELLIGENT ROBOTS AND SYSTEMS, IROS | 0 | 2023 |
| iTutor: A Generative Tutorial System for Teaching the Elders to Use Smartphone Applications | PROCEEDINGS OF THE 36TH ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE & TECHNOLOGY, UIST 2023 | 0 | 2023 |
| Large Language Models are Zero-Shot Reasoners | ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 35, NEURIPS 2022 | 43 | 2022 |
| CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities | PROCEEDINGS OF THE 2022 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI' 22) | 59 | 2022 |
| PromptChainer: Chaining Large Language Model Prompts through Visual Programming | EXTENDED ABSTRACTS OF THE 2022 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI 2022 | 19 | 2022 |
| Making Pre-trained Language Models Better Few-shot Learners | 11TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING | 312 | 2021 |
| Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm | EXTENDED ABSTRACTS OF THE 2021 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI'21) | 125 | 2021 |
| Calibrate Before Use: Improving Few-Shot Performance of Language Models | INTERNATIONAL CONFERENCE ON MACHINE LEARNING | 138 | 2021 |

Između 45 radova navedenih u Tablica 5.1 kao najopširniji se ističe rad od Khan et al. [86] budući da se u njemu analizira prisutnost interakcije s velikim jezičnim modelima u nekoliko različitih domena ljudskih djelatnosti. Naime, u ovom preglednom radu autori analiziraju utjecaj upotrebe generativne umjetne inteligencije na interakciju čovjeka i računala. Navode kako ChatGPT, kao primjer napredne tehnologije UI, omogućuje prirodniju i intuitivniju komunikaciju između ljudi i strojeva te nalazi primjenu u raznim domenama poput korisničke službe, obrazovanja i kreiranja sadržaja. Razvoj velikih jezičnih modela omogućuje ostvarivanje naprednijih oblika interakcije sa sustavima UI. Kao primjer daju *InstructGPT*, jedan od alata temeljenih na GPT modelu, koji unaprjeđuje točnost generiranih odgovora korištenjem ljudskih povratnih informacija, čime se povećava njegova primjenjivost u specifičnim situacijama. Nadalje, navodi se primjena velikih jezičnih modela u inteligentnim vozilima te naglašava mogućnost poboljšanja autonomne vožnje i sustava inteligentnog transporta putem konverzacijske UI. Autori također istražuju koncept nadogradnje ljudsko-strojne interakcije (*HiVeGPT*) koji bi mogao balansirati kontrolu i donošenje odluka između sustava UI i ljudskih vozača, osiguravajući sigurnost i učinkovitost.

U nastavku je preostalih 44 rada grupirano u 5 kategorija s obzirom na različita područja primjene interakcije s velikim jezičnim modelima. Tako su prepoznata sljedeća područja: interakcija čovjeka i računala, oblikovanje upita za LLM-ove, razvoj programske podrške, robotika i obrazovanje.

## Područje interakcije čovjeka i računala (engl. *Human Computer Interaction, HCI*)

U području interakcije s velikim jezičnim modelima ističu se radovi [87-96] koji istražuju kako veliki jezični modeli utječu na našu percepciju, razmišljanje i donošenje odluka.

Salikutluk i suradnici u [87] pišu o interakciji između ljudi i velikih jezičnih modela i naglašavaju važnost dizajniranja sustava UI koji pomažu ljudskom razmišljanju umjesto da automatski obavljaju zadatke. Autori navode da ljudi često zapnu u zadacima procjene kada ne mogu smisliti povezana pitanja koja bi im pomogla u davanju procjene, te naglašavaju potrebu za podrškom sustava UI u ovom području. Iako alati UI nude sugestije slične ljudskim, sudionici su pokazali oprez prema njihovoj upotrebi, stoga se sugerira da sustavi UI trebaju nadopunjavati, a ne oponašati ljudske misaone procese. Ključni zaključak je da najbolji rezultati interakcije i izvedbe nastaju kada se koriste komplementarne snage ljudi i sustava UI.

Zatim, Lee i suautori u [88], identificiraju ključnu potrebu za razumijevanjem generativnih sposobnosti jezičnih modela u dizajnu interakcija. Autori naglašavaju da je analiza velikih skupova podataka o interakcijama jedno od učinkovitih rješenja, jer može obuhvatiti razne kontekste interakcije i podržati različita tumačenja uspješne suradnje. Kao primjer, stvoren je skup podataka CoAuthor, koji sadrži podatke o interakcijama između 63 korisnika i četiri instance GPT-3 modela kroz 1445 sesija.

Pregledni rad autora J. Liu [89] daje osvrt na ChatGPT iz perspektive psihologije i HCI-a. U radu se ističe da ChatGPT omogućuje prirodniju komunikaciju s računalima i omogućuje korisnicima da se izražavaju svakodnevnim jezikom bez potrebe za naprednim programskim vještinama ili složenim naredbama. Ova pristupačnost širi primjenu napredne tehnologije u rješavanju problema među raznovrsnim korisnicima. Sposobnost ChatGPT-a da obavlja širok spektar funkcija omogućuje ljudima da se fokusiraju na kreativnije i strateške zadatke, što ne samo da poboljšava radnu učinkovitost, već bi moglo i transformirati prirodu i zahtjeve poslova, dovodeći do novih trendova i prilika na tržištu rada.

Zatim, rezultati objavljeni u radu [90] od strane Zhou i suradnika upućuju na to da se veliki jezični modeli mogu učinkovito koristiti za razvijanje osnovnih ideja i koncepata za projekt, proizvod ili sustav. Iako nema značajne razlike u kvaliteti konačnih ishoda, uključivanje LLM-ova skraćuje vrijeme završetka zadataka uz smanjenje broja koraka i manje vremena provedenog u kasnijim fazama razvoja. Navodi se da iako su LLM-ovi percipirani kao manje pouzdani, mogu se smatrati potencijalnim suradnicima, dok ljudi trebaju zadržati vodeću ulogu. Ovi nalazi nude HCI zajednici duboko razumijevanje utjecaja LLM-ova na kreativne procese, pružajući vrijedne uvide za dizajn budućih interakcija s ovim modelima.

Nadalje, Casheekar i suradnici u preglednom radu [91] pružaju detaljan pregled povijesti i pozadine *chatbot*-ova, uključujući popularne skupove podataka korištene za njihovu obuku, analizu korisničkog sučelja i korisničkog iskustva te propise vezane uz podatke koji se koriste obuci *chatbot*-ova. Kroz analizu prethodnih pregleda literature, identificirani su značajni istraživački nedostaci, osobito u područjima analize oblikovanja korisničkog sučelja i korisničkog iskustva. Članak ističe kako će budućnost *chatbot*-ova ovisiti o napretku u konverzacijskoj umjetnoj inteligenciji, uključujući poboljšano razumijevanje prirodnog jezika i emocionalnu inteligenciju, što može značajno unaprijediti interakciju između čovjeka i računala.

Koncept *Multimodal Knowledge Reasoning via Chain-of-Thought* (MuKCoT) za zadatke odgovaranja na pitanja vizualnog tipa koja zahtijevaju vanjsko znanje (engl. *Knowledge -based visual question answering*, KBVQA) predstavljen je u radu [92], (Qiu i suradnici). Ovakvi tipovi zadataka zahtijevaju razumijevanje i zaključivanje kroz više modaliteta. Autori tvrde da je uključivanje lanaca zaključivanja obogaćenih multimodalnim znanjem ključno za ove zadatke, stoga koncept MuKCoT koristi sposobnost lančanog razmišljanja velikih jezičnih modela, obogaćenu vizualno utemeljenim znanjem.

Zatim, u radu [93] Tang i suautori predstavljaju *IPSeg* okvir za segmentaciju okoline koristeći vizualne koncepte iz jedne slike. Predloženi *IPSeg* okvir uklanja potrebu za iscrpnim uvježbavanjem modela, nudeći učinkovito i skalabilno rješenje za segmentaciju. *IPSeg* koristi jednu sliku koja sadrži subjektivni vizualni koncept kao fleksibilni upit za vizualne modele poput *DINOv2* i *Stable Diffusion*. Eksperimenti na skupovima podataka kao što su *COCO* i *PASCAL* *VOC* pokazuju učinkovitost *IPSeg*-a, prikazujući značajna poboljšanja u kvaliteti segmentacije i sposobnosti generalizacije.

Procko, Elvira i Ochoa u članku [94] ističu važnost novog praktičnog konteksta korištenja velikih jezičnih modela za generiranje sadržaja poput tekstova, slika i videa. Ovaj pristup omogućuje korisnicima da bez dubokog teorijskog ili tehničkog znanja o umjetnoj inteligenciji prosperiraju jednostavnim davanjem uputa, čime se otvara mogućnost interakcije s općim inteligentnim agentima u prirodnom jeziku. Naglašava se potreba za sviješću o značajnim utjecajima generativne umjetne inteligencije u velikom opsegu, te važnost definiranja njezinih ograničenja kako bi se omogućila simbioza između ljudi i računala. Ova sinergija može ubrzati razmišljanje, učenje, istraživanje i napredak u mnogim primjenama.

Još jedna primjena ChatGPT alata u području HCI-a može se naći u članku [95]. Autori Gao i suradnici navode da je ChatGPT izvrstan alat za interakciju sa inteligentnim transportnim sustavima jer omogućuje komuniciranje sa strojevima na prirodniji i intuitivniji način uz pomoć raznih senzora i multimodalnih tehnologija. Međutim, još uvijek postoje ograničenja ChatGPT algoritama i prepreke hardverskih uređaja koje je potrebno prevladati kako bi se postigla praktična primjena. Jedan od problema je sklonost ChatGPT-a prema davanju dugih i detaljnih odgovora, što nije uvijek u skladu s potrebama vozača koji zahtijevaju kratke i sažete informacije za brze odluke tijekom vožnje. Dulji odgovori mogu dovesti do propuštanja ključnih uputa ili detalja. Također, iako ChatGPT-ovi odgovori mogu djelovati uvjerljivo, nisu uvijek točni te mogu sadržavati pogrešne ili besmislene informacije. Prihvaćanje takvih netočnih odgovora bez ljudske procjene može uzrokovati ozbiljne nesreće. Da bi se zadovoljile stvarne potrebe za točnim, pravovremenim i učinkovitim informacijama, multimodalni sustavi trebaju snažne sposobnosti obrade teksta, glasa i slike.

Konačno, Shu i ostali u [96] opisuju okvir za izradu sustava za preporuke RAH (*RecSys-Assistant-Human*), koji uključuje korištenje LLM agenata. Provedeni eksperimenti pokazuju učinkovitost *Learn-Act-Critic* petlje i mehanizama povratne informacije u prilagođavanju asistenta osobnostima korisnika.

## Područje oblikovanja upita za velike jezične modele (engl. *LLM Prompting*)

Sljedeće područje koje je identificirano kao relevantno je područje oblikovanja upita za velike jezične modele i analizirana su ukupno 22 rada, [97-118].

U preglednom radu [97], Al-Hasan i suautori uspostavljaju referentnu točku za procjenu učinkovitosti ChatGPT-a u zadacima davanja preporuka, uspoređujući njegovu izvedbu s konvencionalnim modelima preporuka. Proces korištenja ChatGPT-a za zadatke davanja preporuka sastoji se od izrade prilagođenih upita prema specifičnostima zadatka, unosa tih upita u ChatGPT model te generiranja preporuka u skladu s navedenim smjernicama, nakon čega slijedi ispitivanje i poboljšanje generiranih preporuka. U inženjerstvu oblikovanja upita za zadatke davanja preporuka, pažljivo se oblikuju upiti kako bi se dobili relevantni rezultati koji zadovoljavaju korisničke preferencije i kontekst, čime se poboljšava relevantnost i personalizacija preporuka temeljenih na ChatGPT-u.

Prema Chenu i suautorima u [98] veliki jezični modeli mogu se koristiti kao temelj sustava za davanje preporuka korisnicima, ponajprije iz razloga što su uvježbani na velikoj količini podataka kako bi učinkovito obrađivali ljudski jezik, uključujući kontekst, metafore i složenu sintaksu. To im omogućuje kreiranje prilagođenih objašnjenja koja su precizna, prirodna i prilagodljiva različitim korisničkim preferencijama. Osim toga, jedinstvene sposobnosti učenja unutar konteksta, poput zero-shot, few-shot i chain-of-thought tehnika, omogućuju LLM-ovima prikupljanje povratnih informacija korisnika u stvarnom vremenu tijekom interakcija.

Zatim, u radu [99] Parker, Anderson, Stone i Oh istražuju kako chain-of-thought tehnika postavljanja upita može povećati točnost velikih jezičnih modela, ali i pružiti uvid u njihov način razmišljanja. Navode da, iako postoji mogućnost da LLM samo imitira racionalno razmišljanje umjesto da pokaže stvarni proces donošenja odluka, zaključci modela su logični i dosljedni s rezultatima, uključujući elemente uzročno-posljedičnog razmišljanja. Također, promjene u upitima su na odgovarajući način mijenjale rezultate modela.

CoT tehniku u kontekstu postavljanja upita istraživali su i Zhang i suradnici u radu [100]. Naime, predstavili su novi pristup za sažimanje više dokumenata pod nazivom *Attend-Arrange-Abstract Chain-of-Thought* (3A-COT). Ova metoda uključuje tri ključne faze: "*Attend*" (izdvajanje ključnih elemenata), "*Arrange*" (organizacija odnosa između elemenata) i "*Abstract*" (generiranje sažetka). Metoda 3A-COT pokazala je bolje rezultate u odnosu na druge metode temeljene na velikim jezičnim modelima na četiri testna skupa: *Multi-News*, *Human-edited Multi-News*, *WCEP-10* i *Multi-XScience*. Evaluacija putem *ROUGE* i *BERTScore* mjerila pokazala je značajna poboljšanja u kvaliteti generiranih sažetaka.

Nadalje, Gorer i Aydemir u [101] opisuju tehnike učenja upita koje su se pokazale obećavajućim pristupom za automatsku generaciju i prilagodbu upita za LLM-ove. Iako su tehnike kao što su *few-shot* i *CoT* pokazale uspjeh, one se oslanjaju na ručno kreiranje i prilagodbu upita, što predstavlja izazov. Automatskim generiranjem smanjuje se potreba za ručnim stvaranjem i prilagodbom upita za nove zadatke.

U radu [102] Ngu, Lee i Shakarian ispituju mjere raznolikosti kao neovisne pokazatelje vjerojatnosti neuspjeha za zadatke postavljanja upita jezičnim modelima. Formalizirane su tri mjere raznolikosti vezane uz postavljanje upita jezičnim modelima koje ne zahtijevaju pristup implementaciji modela ili posjedovanje domenskog znanja. Predstavljen je i koncept upotrebe mjere raznolikosti u kombinaciji s tehnikom postavljanja upita CoT. Eksperimentalni rezultati pokazuju da ova kombinacija smanjuje vjerojatnost neuspjeha za 30% u usporedbi s korištenjem samo jedne tehnike, a u 18 od 20 upita postignuto je poboljšanje.

Prema istraživanju Koime i ostalih u [103], metoda *Zero-shot-CoT* omogućava velikim jezičnim modelima izvođenje složenih „razmišljanja“ bez potrebe za primjerima, koristeći jednostavan upit poput "Razmislimo korak po korak." Ova metoda značajno poboljšava performanse u raznim zadacima, uključujući aritmetiku i simboličko razmišljanje, u usporedbi sa standardnim zero-shot metodama. Metoda zero-shot-CoT pokazuje veliku svestranost i zahtijeva minimalnu prilagodbu upita, što je čini učinkovitijim rješenjem u odnosu na pristupe koji zahtijevaju specifične primjere za svaki zadatak.

Potpuno novu paradigmu koja koristi heuristiku pri oblikovanju upita za LLM uvode Shao i suradnici u članku [104] i to u svrhu poboljšanja kapaciteta LLM-ova u učenju iz malo primjera (*few-shot*). Opsežna ispitivanja, komparativni eksperimenti i sveobuhvatne analize na dva zahtjevna skupa podataka pokazuju superiornost ove metode, nazvane *Prophet Framework*, u odnosu na sve postojeće najnovije metode.

Nadalje, u području primjene LLM-ova i učenja iz malo primjera Gao, Fisch i Chen u [105] predstavili su novu metoda koja poboljšava *few-shot* učenje za manje jezične modele. Pristup nazvan LM-BFF *(engl. better few-shot fine-tuning of language models)* kombinira *fine-tuning* baziran na upitima i automatizirano generiranje upita, uz unaprijeđenu strategiju za uključivanje primjera. Metoda je testirana na različitim NLP zadacima, pokazujući značajna poboljšanja do 30% u odnosu na standardni *fine-tuning*. LM-BFF metoda je posebno učinkovita za zadatke koji se mogu formulirati kao "popuni prazninu", imaju kratke ulazne sekvence i mali broj izlaznih klasa. Iako postoje određena ograničenja za strukturirane zadatke predviđanja, ovaj pristup je općenito primjenjiv na različite zadatke i učinkovit za primjenu u stvarnom svijetu s ograničenim računalnim resursima.

S druge strane, Reynolds i McDonell u radu [106], uvode koncept programiranja upita za velike jezične modele, koji nadilazi paradigmu *few-shot* učenja. Autori objašnjavaju da se oblikovanje upita može shvatiti kao programiranje na prirodnom jeziku, pri čemu se ističu posebnosti GPT-3 modela poput njegove sličnosti superpoziciji autora umjesto jednom ljudskom autoru. Rad predstavlja inovativne pristupe poput subtraktivnog pristupa programiranju upita i tehnika za prevladavanje ograničenja modela u predviđanju dinamike složenijih situacija. Uvodi se pojam metaupita kao fleksibilnog alata za generiranje upita za specifične zadatke, što potencijalno rješava problem ograničene učinkovitosti općenitih upita. Autori također naglašavaju potrebu za novim metodama evaluacije koje bi bolje ocjenjivale sposobnosti velikih jezičnih modela, uključujući izolaciju katastrofalnih pogrešaka i korištenje metaupita u procjenama.

Međutim, Zhao i suautori u svojoj studiji [107] pokazuju da je izvedba GPT-3 modela u *few-shot* učenju vrlo nestabilna i značajno varira ovisno o formatu upita, odabranim primjerima za treniranje i njihovom redoslijedu. Autori su identificirali da ta nestabilnost proizlazi iz pristranosti u jezičnim modelima, poput sklonosti prema češćim oznakama, recentnijim ili uobičajenijim tokenima. Kako bi riješili ovaj problem, razvili su metodu kontekstualne kalibracije koja prilagođava izlazne vjerojatnosti modela, čime se poboljšava točnost, smanjuje varijanca i općenito povećava učinkovitost alata poput GPT-3 za krajnje korisnike. Ovo istraživanje naglašava važnost pažljivog pristupa *few-shot* učenju u obradi prirodnog jezika te otvara nove smjerove za buduća istraživanja u području razvoja robusnijih metoda i dubljeg razumijevanja načina na koji GPT-3 uči iz zadanih upita.

Zatim, u radu [108] Quan i Chen istražuju kako ChatGPT reagira na pitanja različitih razina pristojnosti. Autori pokazuju da model pruža kraće i manje pozitivne odgovore u slučaju kad su postavljena pitanja manje pristojna, te da se model čini manje responzivan kada se suoči s uvredljivim upitima.

Zanimljivo istraživanje proveo je Chang u [109] koristeći Sokratovu metodu za oblikovanje predložaka upita za interakciju s velikim jezičnim modelima kako bi se poboljšala kvaliteta i dosljednost njihovih rezultata. U radu su istaknute ključne tehnike koje usmjeravaju generirane odgovore modela. Kroz primjere je prikazana sposobnost ovih metoda u pojašnjavanju korisničkih upita i ocjenjivanju kvalitete generiranog teksta.

Nadalje, u članku [110], Kochanek i ostali ističu da iako sinteza podataka temeljenih na upitima za LLM-ove ne može u potpunosti zamijeniti standardno prikupljanje podataka, vrlo je korisna u okruženjima s ograničenim podacima i kao tehnika augmentacije podataka. Pokazalo se da upiti daju obećavajuće rezultate u zadacima poput ekstrakcije odnosa, prikaza lanca razmišljanja i sintetičkih dijaloga. Iako predložene metode još nisu dovoljno napredne da bi postigle rezultate usporedive s punim skupovima podataka, uvijek donose poboljšanja u usporedbi s izvornim nebalansiranim skupovima podataka.

Prema Ma i suradnicima u [111], modeli poput *GPT-3* i *BLOOM* pokazali su značajne sposobnosti u učenju iz konteksta, međutim rezultati mogu varirati zbog izabranih primjerima i formata upita koji se zadaju. Stoga je konstrukcija odgovarajućeg upita ključna za poboljšanje performansi. Predložene su strategije *T-fair-Prompting* i *G-fair-Prompting* za optimizaciju upita smanjenjem prediktivne pristranosti, pri čemu *G-fair-Prompting* nudi bolju ravnotežu između kvalitete pretraživanja i troškova, što je potvrđeno eksperimentalnim rezultatima s modelima kao što su *GPT-3* i *LLaMA*.

Cheng i suradnici su u [112] predstavili metodologiju za oblikovanje sustava UI koji se sastoje od različitih modela ili komponenti koji surađuju kako bi postigli određeni cilj. Metodologija obuhvaća tri ključna koncepta i četiri aktivnosti, s ciljem poboljšanja modularnosti, sastavljivosti, otklanjanja grešaka i ponovne upotrebljivosti funkcionalnosti sustava. Kroz ovu metodologiju uveden je *Prompt Sapper*, vizualni programski alat koji omogućava sastavljanje sustava od osnovnih modela pomoću oblikovanja upita i vizualnog programiranja.

Nadalje, Park i ostali autori u [113], predlažu četiri tehnike postavljanja upita: opći upit, upit s prethodnim frazama, *cloze* upit i upit s eksplicitnom svrhom. Tehnika upita s prethodnim frazama (PPQ) uključuje davanje početne rečenice ili riječi potrebnih za relevantan i prikladan odgovor, omogućujući modelu generiranje sljedeće rečenice. *Cloze* upit (CQ) tehnika podrazumijeva prikaz rečenice s praznim mjestima koja se popunjavaju primjerima, što je često korišteno u modelima generiranja jezika. Upit s eksplicitnom svrhom (PEQ) uključuje navođenje svrhe ili cilja te pružanje kategoriziranih primjera u skladu s tim, s ciljem da modeli generiranja jezika proizvedu prirodnije odgovore razumijevanjem specifične svrhe upita.

Nova tehnika oblikovanja upita za LLM-ove, predstavljena u [114] (Seo, Zhang i Rong), poboljšava sposobnost velikih jezičnih modela za prepoznavanje zlonamjernog koda, čime se pomaže u osiguravanju sigurnosti sustava i podataka.

Kao što je navedeno u radu [115] (Wu et al.), koncept povezivanja više upita za velike jezične modele s ciljem upravljanja složenim zadacima, koji ne mogu biti riješeni jednim izvođenjem LLM-a, predstavlja ključnu temu. U radu je predstavljen alat *PromptChainer*, interaktivno sučelje za vizualno programiranje lanaca LLM-ova, koje omogućuje korisnicima transformaciju podataka između koraka i ispravljanje grešaka na različitim razinama detalja. Autori identificiraju potrebe korisnika pri stvaranju lanaca LLM-ova, uključujući podršku za transformaciju podataka, otklanjanje pogrešaka i izgradnju mentalnog modela sposobnosti LLM-a. Također, rad prikazuje kako *PromptChainer* podržava izgradnju prototipova za različite primjene te raspravlja o budućim smjerovima za skaliranje i poboljšanje lanaca.

Wang, Li i Li u radu [116], primijenili su velike jezične modele za omogućavanje razgovornih interakcija na mobilnim korisničkim sučeljima. Cilj studije bio je istražiti mogućnost korištenja jednog LLM-a za raznovrsne razgovorne interakcije s mobilnim UI-ima bez potrebe za specifičnim skupovima podataka i obukom za svaku zadaću. Autori su razvili tehnike upita kako bi prilagodili LLM mobilnim UI-ima i testirali ih na četiri zadatka: generiranje pitanja za trenutni sadržaj zaslona, sažimanje sadržaja zaslona, odgovaranje na pitanja sa zaslona i mapiranje uputa u akcije na sučelju. Rezultati su pokazali konkurentne performanse, što ukazuje na potencijal LLM-a za podršku raznovrsnim razgovornim interakcijama s mobilnim sučeljima.

Istraživanje koje su proveli Zamfrescu-Pereira i ostali suradnici u [117], pokazuje da se mnogi korisnici koji nisu stručnjaci u području UI suočavaju s izazovima pri oblikovanju učinkovitih upita za velike jezične modele zbog sklonosti oslanjanju na svoja iskustva iz međuljudske komunikacije i sklonosti prekomjernom generaliziranju. Studija je koristila prototipni alat za dizajn *chatbot*-a temeljen na LLM-u kako bi istražila pristup ne-stručnjaka u oblikovanju upita. Sudionici su pristupili oblikovanju upita oportunistički, a ne sustavno. Rezultati sugeriraju potrebu za boljim alatima za oblikovanje i obrazovanjem kako bi se poboljšala pismenost o LLM-ovima i upitima među ne-stručnjacima, što otvara mogućnosti za daljnja istraživanja.

I na kraju, Bridgelall u [118] objašnjava kako je inženjerstvo oblikovanja upita postalo nova vještina potrebna kako bi se dobili željeni rezultati od velikih jezičnih modela. Navodi kako se dobar upit sastoji se od četiri dijela: postavljanja konteksta, specifične upute, željenog formata/strukture i opcionalno primjera. Postavljanje konteksta daje kratak pregled pozadine teme ili zadatka. Specifične upute jasno navode zahtjeve, pitanja ili zadatke, smanjujući šanse za generiranje nevažnih odgovora. Format/struktura specificira kako trebaju izgledati generirani odgovori, uključujući broj riječi, oznake i naslove. Primjeri pomažu LLM-u da bolje razumije željeni stil, format ili izlaz, a rezultati se poboljšavaju s jednim ili više primjera. U radu se još navodi i da su sadašnja istraživanja usmjerena unapređivanju multimodalnih modela i onih modela koji zahtijevaju manje resursa, dok u budućnosti predviđaju razvoj usmjeren postizanju generalne umjetne inteligencije.

## Područje razvoja programske podrške (engl. *SW development*)

Treće područje je primjena velikih jezičnih modela u kontekstu razvoja programske podrške. Iako su u ovom pregledu identificirana samo dva rada koja se izravno bave ovom tematikom [119, 120], područje zaslužuje pozornost zbog svog rastućeg značaja i potencijala. LLM-ovi već sada pokazuju izvanredne sposobnosti u zadacima kao što su generiranje koda, otklanjanje pogrešaka i automatizacija rutinskih programerskih zadataka. Štoviše, integracija LLM-ova u razvojna okruženja i alate za programiranje potencijalno najavljuje novu eru u programskom inženjerstvu, gdje bi suradnja između ljudskih programera i asistenata UI mogla značajno ubrzati i unaprijediti proces razvoja programske podrške.

Tako, Munley, Jarmusch i Chandrasekaran u svom radu [119] istražuju automatsku generaciju testova za validaciju i verifikaciju programskog koda korištenjem velikih jezičnih modela. Ispitani su modeli otvorenog koda poput *Codellama* i *Deepseek Coder* modela, te modeli zatvorenog koda kao što su *GPT-3.5-Turbo* i *GPT-4-Turbo*. Primijenjene su tehnike postavljanja upita koje uključuju predloške koda, *Retrieval Augmented Generation* (RAG) tehniku. Najbolje rezultate u generiranju uspješnih testova postigao je *Deepseek-Coder-33b-Instruct*, dok je *GPT-4-Turbo* bio sljedeći po uspješnosti.

S druge strane, Marques, Silva i Bernardino u preglednom radu [120] predstavili su skup predložaka za oblikovanje upita za LLM-ove koji se mogu koristiti u raznim zadacima programskog inženjerstva, kao što je npr. definiranje zahtjeva. Skup predložaka za definiranje zahtjeva uključuje elemente kao što su razjašnjavanje specifikacija ili simulacija zahtjeva za promjenom. Premda rad pruža sistematsku klasifikaciju i tehnike oblikovanja upita za unapređenje različitih aspekata programskog inženjerstva, nedostaje mu dubinska analiza praktične primjene ovih predložaka i ocjena njihove učinkovitosti.

## Područje robotike (engl. *Robotics*)

U području robotike iskazan je poseban interes za integraciju velikih jezičnih modela u fizičke sustave s ciljem poboljšanja interakcije između ljudi i robota. Inženjerski aspekti usmjereni su na praktičnu primjenu ovih modela u različitim industrijskim i poslovnim kontekstima. U području robotike identificirana su 4 rada, [121-124].

Vemprala i suradnici u radu [121] prikazuju kako se ChatGPT može koristiti za rješavanje različitih zadataka u području robotike, od osnovnog logičkog, geometrijskog i matematičkog zaključivanja do složenih domena kao što su zračna navigacija, manipulacija i utjelovljeni agenti. Osmišljen je istraživački alat otvorenog koda pod nazivom *PromptCraft* koji omogućuje istraživačima da zajednički dijele i ocjenjuju primjere dobrih shema upita za LLM-ove koji se mogu primijeniti u radu s robotima te nudi simulator robotike s integracijom ChatGPT-a. U radu se također opisuje kako korisnici mogu koristiti različite tehnike upita, uključujući dijalog u slobodnom obliku, kodiranje upita, XML oznake i zatvorenu petlju zaključivanja, kako bi model mogao brzo analizirati ljudsku namjeru i generirati kod za rješavanje problema. U ovom radu ispituje se ideja korištenja ChatGPT-a tijekom procesa razvoja robotskih algoritama, ali buduća istraživanja mogla bi ići u smjeru korištenja LLM-ova u aplikacijama robotike u stvarnom vremenu uz osiguravanje sigurnosti i ispravnosti generiranog koda.

Nadalje, Ye, You i Du navode u radu [122] kako integracija ChatGPT-a u robote značajno povećava povjerenje u suradnji između ljudi i robota zahvaljujući poboljšanoj komunikaciji i razumijevanju. U istraživanju je razvijen sustav RoboGPT koji koristi ChatGPT za upravljanje robotom putem prirodnog jezika. Rezultati eksperimenta s ljudskim sudionicima pokazuju da je sustav poboljšao izvedbu zadataka i smanjio kognitivno opterećenje u usporedbi s fiksnim sustavima za davanje naredbi. Nalazi sugeriraju da uvođenje velikih jezičnih modela poput ChatGPT može unaprijediti učinkovitost i pouzdanost sustava suradnje između ljudi i robota.

S druge strane, Liu i ostali autori u [123], predlažu novu metodu zasnovanu na LLM-ovima za poboljšanje upravljanja robotom. To se postiže kroz suradnju čovjeka i robota integrirajući GPT-4 model s *YOLO* algoritmom za percepciju okoline. Navodi se kako LLM može razložiti složene zadatke na podzadatke i izvršavati ih sekvencijalno, čime se povećava autonomija i učinkovitost robota. Autonomija robota unapređuje se putem sučelja koje povezuje LLM s robotskim sustavima i omogućuje izvršavanje zadataka putem prirodnog jezika.

Konačno, Zhao i suautori napravili su istraživanje na temu interaktivne multimodalne percepcije u [124] i navode da agent *Matcha* koristi velike jezične modele za rasuđivanje na temelju osjetila kao što su vid, zvuk i dodir u svrhu izvršenja zadatka. Arhitektura ovog agenta uključuje LLM kao bazu i module za multimodalnu percepciju, pri čemu se koristi prirodni jezik za razmjenu informacija. Eksperimentalni rezultati pokazuju visoku uspješnost agenta *Matcha* u zadacima prepoznavanja objekata, koja nadmašuje standardne metode. Modularni okvir omogućuje jednostavnu integraciju novih radnji ili percepcijskih modula, što ukazuje na široku primjenjivost u različitim robotskim scenarijima.

## Područje obrazovanja (engl. *Education*)

U ovoj domeni identificirano je 6 radova, [125-130]. Ovi radovi su fokusirani na mogućnosti primjene umjetne inteligencije u unapređenju nastavnih metoda i personalizaciji učenja.

Prvi u nizu je pregledni rad [125] autora Mai, Da i Hanh. Autori navode da je snaga ChatGPT-a u njegovoj sposobnosti interakcije kroz prirodne razgovore, uključujući seriju pitanja od korisnika i povratne informacije od aplikacije. Ističe se da je ključno učenicima omogućiti vještine pisanja upita kako bi se maksimizirale prednosti ChatGPT-a u obrazovanju. Te vještine omogućuju učenicima da učinkovito komuniciraju s modelom, čime se povećava vjerojatnost dobivanja točnih i relevantnih odgovora. ChatGPT također pomaže nastavnicima u njihovim zadacima podučavanja, omogućujući personaliziraniju interakciju sa studentima, potiče značajne dijaloge i aktivno sudjelovanje te pruža povratne informacije i ocjene u stvarnom vremenu, čime doprinosi kontinuiranom poboljšanju učenika tijekom njihovog obrazovnog puta.

Zatim, Pelaez-Sanchez, Velarde-Camaqui i Glasserman-Morales u preglednom radu [126] ističu značajan potencijal velikih jezičnih modela za unaprjeđivanje visokog obrazovanja, u skladu s načelima Obrazovanja 4.0, koja promiču autonomno, suradničko i interaktivno učenje. Ovi modeli mogu simulirati interaktivne sesije za poučavanje i pružati objašnjenja prilagođena potrebama svakog pojedinog studenta, čime se poboljšava razumijevanje i zadržavanje znanja. Međutim, integracija umjetne inteligencije u obrazovanje zahtijeva ljudski nadzor kako bi se osigurala kvaliteta, točnost i etička provedba.

Hwang i Chen u svojoj studiji [127] predlažu da istraživači i edukatori zamijene "pretraživački" način razmišljanja s "programiranjem upita" pri radu s aplikacijama generativne umjetne inteligencije (GAI). "Programiranje upita" odnosi se na sposobnost usmjeravanja GAI aplikacija putem niza logičkih uputa za izvršenje zadataka. Dobro osmišljeni upiti mogu voditi ChatGPT do kvalitetnog obavljanja zadatka. Ovo ukazuje na to da bi, iz perspektive obrazovanja, razvijanje kompetencija učitelja i učenika za oblikovanje upita značajno utjecalo na kvalitetu podučavanja i učenja temeljenog na GAI, uključujući kvalitetu nastavnog sadržaja, oblikovanja učenja i ocjenjivanja za učitelje, te ishoda učenja za učenike.

Zanimljivo istraživanje na temu kako djeca povjeravaju informacije sustavima umjetne inteligencije koji su dizajnirani da izgledaju prijateljski i privlačno predstavlja Kurian u [128]. U istraživanju je korišten humanoidni robot NAO poznat znanstvenoj i općoj populaciji po svom simpatičnom izgledu i dječjoj veličini. Studija je pokazala da su djeca otkrila više informacija o svom mentalnom zdravlju robotu nego ljudskom sugovorniku pa su tako autori zaključili da su djeca smatrala robota pouzdanijim u usporedbi s odraslim sugovornicima. Naime, djeca su vjerovala da se neće naći u nevolji nakon što podijele svoje tajne s robotom.

Nadalje, Zou, Ye i Ye u [129] predstavljaju generativni sustav za podučavanje starijih osoba u korištenju pametnih telefona pod nazivom iTutor. Ovaj sustav se ističe svojom sposobnošću dinamičkog generiranja uputa temeljem trenutnih ciljeva i konteksta korisničkog sučelja, što je omogućeno primjenom tehnika oblikovanja upita na velike jezične modele.

Naposljetku, Bozkurt u [130] ističe se važnost uključivanja oblikovanja upita za LLM-ove u obrazovne kurikulume. Podučavanjem učenika kako formulirati jasne i kontekstualno relevantne upite, edukatori mogu osnažiti učenike da učinkovito koriste tehnologije UI za personalizirana iskustva učenja. Učenici koji ovladaju tehnikama postavljanja upita moći će dobiti preciznije i dublje odgovore od sustava UI, čime će poboljšati svoje vještine rješavanja problema i stjecanje znanja. Nadalje, edukatori mogu kreirati upite koji potiču kritičko razmišljanje i rješavanje problema, što vodi do dubljeg angažmana i interaktivnijeg okruženja za učenje. Ovaj pristup omogućava kreiranje obrazovnog sadržaja poboljšanog umjetnom inteligencijom koji se usklađuje sa specifičnim obrazovnim ciljevima i podržava pružanje vrijednih, individualiziranih povratnih informacija. U ovom kontekstu, kako UI nastavlja igrati sve značajniju ulogu u obrazovanju, razvijanje stručnosti u oblikovanju upita postaje sve relevantnija.

# ZAKLJUČAK

U ovom kvalifikacijskom doktorskom ispitu predstavljen je opsežan pregled interakcije sa sustavima umjetne inteligencije temeljenim na velikim jezičnim modelima. Obuhvaćen je širok spektar tema i pristupa kroz sustavnu analizu relevantne literature. U radu su obrađeni ključni aspekti razvoja, primjene i izazova povezanih s ovim naprednim tehnologijama.

Kroz analizu velikih jezičnih modela, istražene su osnovne postavke modela, tehnike postavljanja upita, praktični primjeri korištenja te usporedba performansi modela. Dodatno, prikazan je primjer interakcije s NAO robotom, čime su ilustrirane praktične primjene teorijskih postavki. Konačno, rad se osvrće na interakciju sa sustavima umjetne inteligencije, navodeći taksonomiju i smjernice za oblikovanje interakcije, te istražuje čovjeku usmjerenu umjetnu inteligenciju, agentske sustave i bihevioristički pristup analizi sustava, pružajući opsežan pregled i smjernice za buduće istraživanje u ovom polju.

Pregledom literature pokazano je da se istraživanja u ovom području protežu kroz nekoliko znanstvenih disciplina, uključujući računarstvo, psihologiju, obrazovanje, robotiku i inženjerstvo. Ova interdisciplinarnost odražava široki utjecaj i potencijal velikih jezičnih modela u različitim domenama ljudske djelatnosti.

U svrhu lakšeg razumijevanja identificirano je nekoliko područja primjene. Tako je u području računarstva fokus stavljen na tehnike oblikovanja upita, poboljšanje performansi i sigurnosti modela. S dijelovima koda i čak cijelim programima koji se mogu automatski generirati pomoću velikih jezičnih modela, umijeće oblikovanja upita postaje nova specifikacija zahtjeva. S druge strane, psihološka istraživanja bavila su se pitanjima interakcije između čovjeka i sustava umjetne inteligencije, ispitujući kako ovi sustavi utječu na ljudsku percepciju, razmišljanje i donošenje odluka. U kontekstu obrazovanja, radovi su se fokusirali na potencijal umjetne inteligencije u unapređenju nastavnih metoda i personalizaciji učenja. Konačno, u području robotike javlja se rastuće zanimanje za ugradnju velikih jezičnih modela u fizičke sustave, s ciljem unaprjeđenja interakcije između ljudi i robota.

Analizirani radovi pokazuju da, iako veliki jezični modeli nude izvanredne mogućnosti u procesiranju i generiranju prirodnog jezika i drugog sadržaja, još uvijek postoje značajni izazovi u njihovoj primjeni. Ovi izazovi uključuju etička pitanja, probleme pristranosti, potrebu za boljom kontrolom i razumijevanjem njihovog ponašanja, te optimizaciju njihove primjene u specifičnim domenama.

Stoga se buduće istraživanje može usmjeriti na nekoliko ključnih područja:

* Razvoj robusnijih metoda za evaluaciju i usporedbu velikih jezičnih modela, posebno u kontekstu njihove primjene.
* Istraživanje naprednih tehnika oblikovanja upita koje bi omogućile precizniju i pouzdaniju interakciju s modelima.
* Integracija velikih jezičnih modela s drugim tehnologijama umjetne inteligencije, poput računalnog vida i robotike, za stvaranje sveobuhvatnijih sustava UI.
* Razvoj metoda za poboljšanje transparentnosti i objašnjivosti odluka koje donose veliki jezični modeli, što je ključno za njihovu širu prihvaćenost i primjenu.
* Razvoj etičkih smjernica i regulatornih okvira za primjenu ovih tehnologija, posebno u osjetljivim područjima poput obrazovanja.

Zaključno, ovaj rad pruža temelj za daljnje istraživanje i razvoj u području interakcije sa sustavima umjetne inteligencije temeljenim na velikim jezičnim modelima. Budući razvoj u ovom području ima potencijal unaprijediti način na koji komuniciramo s tehnologijom i oblikujemo naše društvo u cjelini.

# LITERATURA

|  |  |
| --- | --- |
| [1] | C. Zhang i Y. Lu, »Study on artificial intelligence: The state of the art and future prospects,« *Journal of Industrial Information Integration,* svez. 23, p. 100224, 2021. |
| [2] | IBM, »What is artificial intelligence (AI)?,« IBM, [Mrežno]. Available: https://www.ibm.com/topics/artificial-intelligence. [Pokušaj pristupa 11 July 2024]. |
| [3] | D. Stipaničev, L. Šerić i M. Braović, Uvod u umjetnu inteligenciju, Split: Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, 2021. |
| [4] | IBM, »What is a neural network?,« IBM, [Mrežno]. Available: https://www.ibm.com/topics/neural-networks. [Pokušaj pristupa 11 July 2024]. |
| [5] | R. I. Mukhamediev, Y. Popova, Y. Kuchin, E. Zaitseva, A. Kalimoldayev, A. Symagulov, V. Levashenko, F. Abdoldina, V. Gopejenko, K. Yakunin, E. Muhamedijeva i M. Yelis, »Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges,« *Mathematics,* svez. 10, 2022. |
| [6] | NVIDIA, »Generative AI – What is it and How Does it Work?,« NVIDIA , [Mrežno]. Available: https://www.nvidia.com/en-us/glossary/generative-ai/. [Pokušaj pristupa 12 July 2024]. |
| [7] | A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser i I. Polosukhin, »Attention is All you Need,« u *Advances in Neural Information Processing Systems*, 2017. |
| [8] | G. Yenduri et al., »GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions,« *IEEE Access,* svez. 12, pp. 54608-54649, 2024. |
| [9] | M. R. Douglas, »Large Language Models,« *arXiv preprint arXiv:2307.05782,* 2023. |
| [10] | F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau i L. Chen, »Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration,« *Journal of Information Technology Case and Application Research,* svez. 25, br. 3, p. 277–304, 2023. |
| [11] | Team Gemini, et al., »Gemini: a family of highly capable multimodal models,« *arXiv preprint arXiv:2312.11805,* 2023. |
| [12] | L. Li, M. Doroslovački i . H. Loew, »Approximating the Gradient of Cross-Entropy Loss Function,« *IEEE Access,* svez. 8, pp. 111626-111635, 2020. |
| [13] | L. Beurer-Kellner, M. Fischer i M. Vech, »Prompting Is Programming: A Query Language for Large Language Models,« *Proc. ACM Program. Lang. 7, PLDI,* p. 24, 2023. |
| [14] | »Prompt Engineering Guide,« DAIR.AI, [Mrežno]. Available: https://www.promptingguide.ai. [Pokušaj pristupa 29 May 2024]. |
| [15] | »LLM Parameters Demystified: Getting The Best Outputs from Language AI,« Cohere, [Mrežno]. Available: https://cohere.com/blog/llm-parameters-best-outputs-language-ai. [Pokušaj pristupa 29 May 2024]. |
| [16] | »API reference,« OpenAI, [Mrežno]. Available: https://platform.openai.com/docs/api-reference/chat/create. [Pokušaj pristupa 29 May 2024]. |
| [17] | M. McTear i M. Ashurkina, Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents, Apress, 2024. |
| [18] | A. Gao, »Prompt Engineering for Large Language Models,« SSRN, 8 July 2023. [Mrežno]. Available: https://ssrn.com/abstract=4504303. [Pokušaj pristupa 31 May 2024]. |
| [19] | »Prompt engineering,« OpenAI, [Mrežno]. Available: https://platform.openai.com/docs/guides/prompt-engineering. [Pokušaj pristupa 31 May 2024]. |
| [20] | E. van Dis, J. Bollen, W. Zuidema, R. van Rooij i C. Bockting, »ChatGPT: five priorities for research,« *Nature,* svez. 614, pp. 224-226., 2023. |
| [21] | »LLM prompting guide,« Hugging Face , [Mrežno]. Available: https://huggingface.co/docs/transformers/main/tasks/prompting. [Pokušaj pristupa 31 May 2024]. |
| [22] | Hao, Shibo et al., »Reasoning with Language Model is Planning with World Model,« *ArXiv pre-print ArXiv abs/2305.14992,* 2023. |
| [23] | J. White, et al., »A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,« *ArXiv pre-print, ArXiv abs/2302.11382,* 2023. |
| [24] | Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan i J. Ba, »Large Language Models Are Human-Level Prompt Engineers,« *ArXiv,* svez. abs/2211.01910, 2022. |
| [25] | J. Zelle, Python Programming: An Introduction to Computer Science 2nd Edition, USA: Franklin, Beedle & Associates Inc., 2010. |
| [26] | J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai i Q. V. Le, »Finetuned Language Models Are Zero-Shot Learners,« *ArXiv,* svez. abs/2109.01652, 2021. |
| [27] | S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi i L. Zettlemoyer, »Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?,« *ArXiv,* svez. abs/2202.12837, 2022. |
| [28] | T. Brown, et al., »Language Models are Few-Shot Learners,« u *Advances in Neural Information Processing Systems*, 2020. |
| [29] | J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H.-h. Chi, F. Xia, Q. Le i D. Zhou, »Chain of Thought Prompting Elicits Reasoning in Large Language Models,« *ArXiv,* svez. abs/2201.11903, 2022. |
| [30] | X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H.-h. Chi i D. Zhou, »Self-Consistency Improves Chain of Thought Reasoning in Language Models,« *ArXiv,* svez. abs/2203.11171, 2022. |
| [31] | »Chain prompts,« Anthropic , [Mrežno]. Available: https://docs.anthropic.com/en/docs/chain-prompts. [Pokušaj pristupa 1 June 2024]. |
| [32] | S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao i K. Narasimhan, »Tree of Thoughts: Deliberate Problem Solving with Large Language Models,« *ArXiv,* svez. abs/2305.10601, 2023. |
| [33] | J. Long, »Large Language Model Guided Tree-of-Thought,« *ArXiv,* svez. abs/2305.08291, 2023. |
| [34] | P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel i D. Kiela, »Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,« *ArXiv,* svez. abs/2005.11401, 2020. |
| [35] | Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang i H. Wang, »Retrieval-Augmented Generation for Large Language Models: A Survey,« *ArXiv,* svez. abs/2312.10997, 2023. |
| [36] | B. Paranjape, S. M. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer i M. T. Ribeiro, »ART: Automatic multi-step reasoning and tool-use for large language models,« *ArXiv,* svez. abs/2303.09014, 2023. |
| [37] | S. Diao, P. Wang, Y. Lin i T. Zhang, »Active Prompting with Chain-of-Thought for Large Language Models,« *ArXiv,* svez. abs/2302.12246, 2023. |
| [38] | L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan i G. Neubig, »PAL: Program-aided Language Models,« *ArXiv,* svez. abs/2211.10435, 2022. |
| [39] | S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan i Y. Cao, »React: Synergizing reasoning and acting in language models,« *arXiv preprint arXiv:2210.03629,* 2022. |
| [40] | N. Shinn, F. Cassano, B. Labash, A. Gopinath, K. Narasimhan i S. Yao, »Reflexion: language agents with verbal reinforcement learning,« u *Neural Information Processing Systems*, 2023. |
| [41] | »Reflexion: An Iterative Approach to LLM Problem-Solving,« Prompt Engineering Institute, 9 May 2023. [Mrežno]. Available: https://promptengineering.org/reflexion-an-iterative-approach-to-llm-problem-solving/. [Pokušaj pristupa 1 June 2024]. |
| [42] | Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis i A. J. Smola, »Multimodal Chain-of-Thought Reasoning in Language Models,« *ArXiv,* svez. abs/2302.00923, 2023. |
| [43] | D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. X. Song i J. Steinhardt, »Measuring Massive Multitask Language Understanding,« *ArXiv,* svez. abs/2009.03300, 2020. |
| [44] | »Deciphering AI Prowess: A Deep Dive into Language Model Benchmarks,« LinkedIn, 12 December 2023. [Mrežno]. Available: https://www.linkedin.com/pulse/deciphering-ai-prowess-deep-dive-language-model-zou-msc-ma-bsc--ea4jc/. [Pokušaj pristupa 1 June 2024]. |
| [45] | M. Chen, et al., »Evaluating Large Language Models Trained on Code,« *ArXiv,* svez. abs/2107.03374, 2021. |
| [46] | A. Srivastava, »Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models,« *ArXiv,* svez. abs/2206.04615, 2022. |
| [47] | D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh i M. Gardner, »DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs,« u *North American Chapter of the Association for Computational Linguistics*, 2019. |
| [48] | »ARC: AI2 Reasoning Challenge Leaderboard,« Allen Institute for AI , [Mrežno]. Available: https://leaderboard.allenai.org/arc/submissions/about. [Pokušaj pristupa 1 June 2024]. |
| [49] | K. Sakaguchi, R. L. Bras, C. Bhagavatula i Y. Choi, »WinoGrande,« *Communications of the ACM,* svez. 64, pp. 99-106, 2019. |
| [50] | D. Zan, B. Chen, F. Zhang, D. Lu, B. Wu, B. Guan, Y. Wang i J.-G. Lou, »Large Language Models Meet NL2Code: A Survey,« u *Annual Meeting of the Association for Computational Linguistics*, 2022. |
| [51] | »Metrics,« Machine Translate, [Mrežno]. Available: https://machinetranslate.org/metrics. [Pokušaj pristupa 1 June 2024]. |
| [52] | D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael i S. R. Bowman, »GPQA: A Graduate-Level Google-Proof Q&A Benchmark,« *ArXiv,* svez. abs/2311.12022, 2023. |
| [53] | F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das i J. Wei, »Language Models are Multilingual Chain-of-Thought Reasoners,« *ArXiv,* svez. abs/2210.03057, 2022. |
| [54] | OpenAI, »GPT-4 Technical Report,« OpenAI, [Mrežno]. Available: https://cdn.openai.com/papers/gpt-4.pdf. [Pokušaj pristupa 02 June 2024]. |
| [55] | Anthropic, »The Claude 3 Model Family: Opus, Sonnet, Haiku,« [Mrežno]. Available: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model\_Card\_Claude\_3.pdf. [Pokušaj pristupa 2 June 2024]. |
| [56] | »Create a Message,« Anthropic, [Mrežno]. Available: https://docs.anthropic.com/en/api/messages. [Pokušaj pristupa 2 June 2024]. |
| [57] | G. Gemini Team, »Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,« [Mrežno]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini\_v1\_5\_report.pdf. [Pokušaj pristupa 2 June 2024]. |
| [58] | H. Z. Munmun De Choudhury i D. A. Shamma, »Introduction to this special issue on unifying human computer interaction and artificial intelligence,« *Human–Computer Interaction,* svez. 35, p. 355–361, 2020. |
| [59] | A. Granić, »Technology in use: The importance of good interface design,« u *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, 2017. |
| [60] | S. Hinsen, P. Hofmann, J. Jöhnk i N. Urbach, »How Can Organizations Design Purposeful Human-AI Interactions: A Practical Perspective From Existing Use Cases and Interviews.,« u *Proceedings of the Annual Hawaii International Conference on System Sciences*, Kauai, USA, 2022. |
| [61] | M. K. Lee, N. Grgić-Hlača, M. C. Tschantz, R. Binns, A. Weller, M. Carney i K. Inkpen, »Human-Centered Approaches to Fair and Responsible AI,« u *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2020. |
| [62] | »“What is Human-AI Interaction (HAX)?”,« Interaction Design Foundation - IxDF, [Mrežno]. Available: https://www.interaction-design.org/literature/topics/human-ai-interaction . [Pokušaj pristupa 6 June 2024]. |
| [63] | A. Granić i F. D. Davis, The Technology Acceptance Model: 30 Years of TAM (Human-computer interaction series), Springer Nature, 2024. |
| [64] | M. Theofanos, Y. Choong i T. Jensen, »AI Use Taxonomy: A Human-Centered Approach,« National Institute of Standards and Technology, NIST Trustworthy and Responsible AI, Gaithersburg, MD, 2024. |
| [65] | W. Geyer, J. Weisz, C. Santos Pinhanez i E. Daly, »What is human-centered AI?,« IBM, 31 March 2022. [Mrežno]. Available: https://research.ibm.com/blog/what-is-human-centered-ai. [Pokušaj pristupa 14 July 2024]. |
| [66] | »“What is Human-Centered AI (HCAI)?,« Interaction Design Foundation - IxDF, [Mrežno]. Available: https://www.interaction-design.org/literature/topics/human-centered-ai. [Pokušaj pristupa 14 July 2024]. |
| [67] | »Human-Centered AI,« National Institute of Standards and Technology (NIST), 23 April 2024. [Mrežno]. Available: https://www.nist.gov/programs-projects/human-centered-ai. [Pokušaj pristupa 14 July 2024]. |
| [68] | T. Capel i M. Brereton, »What is Human-Centered about Human-Centered AI? A Map of the Research Landscape,« u *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. |
| [69] | Abedin, B.; Meske, C.; Junglas, I.;, »Designing and Managing Human-AI Interactions,« *Information Systems Frontiers,* svez. 24, br. https://doi.org/10.1007/s10796-022-10313-1, p. 691–697, 2022. |
| [70] | D. Shin i E. Y. Shin, »Human-Centered AI: A Framework for Green and Sustainable AI,« *Computer,* svez. 56, pp. 16-25, 2023. |
| [71] | S. Amershi, M. Vorvoreanu i E. Horvitz, »Guidelines for human-AI interaction design,« Microsoft, 1 February 2019. [Mrežno]. Available: https://www.microsoft.com/en-us/research/blog/guidelines-for-human-ai-interaction-design/. [Pokušaj pristupa 17 July 2024]. |
| [72] | »Guidelines for Human-AI Interaction,« Microsoft, [Mrežno]. Available: https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/. [Pokušaj pristupa 17 July 2024]. |
| [73] | S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil i E. Horvitz, »Guidelines for Human-AI Interaction,« u *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2019. |
| [74] | C. Wang, »Agents in AutoGen,« Microsoft Research, 24 May 2024. [Mrežno]. Available: https://microsoft.github.io/autogen/blog/2024/05/24/Agent/. [Pokušaj pristupa 18 July 2024]. |
| [75] | »Getting Started,« AutoGen, [Mrežno]. Available: https://microsoft.github.io/autogen/docs/Getting-Started. [Pokušaj pristupa 18 July 2024]. |
| [76] | »What are LLM Agents?,« Iguazio, [Mrežno]. Available: https://www.iguazio.com/glossary/llm-agents/. [Pokušaj pristupa 18 July 2024]. |
| [77] | S. Lacasse, »Building a multimodal multi-agent framework with Azure OpenAI Assistant API,« Microsoft, 19 March 2024. [Mrežno]. Available: https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/building-a-multimodal-multi-agent-framework-with-azure-openai/ba-p/4084007. [Pokušaj pristupa 18 July 2024]. |
| [78] | J. Zhang, »EcoAssistant - Using LLM Assistants More Accurately and Affordably,« AutoGen, 9 November 2023. [Mrežno]. Available: https://microsoft.github.io/autogen/blog/2023/11/09/EcoAssistant/. [Pokušaj pristupa 18 July 2024]. |
| [79] | J. Zhang, R. Krishna, A. H. Awadallah i C. Wang, »Ecoassistant: Using llm assistant more affordably and accurately,« *arXiv preprint arXiv:2310.03046,* 2023. |
| [80] | W. Tao, Y. Zhou, W. Zhang i Y. Cheng, »MAGIS: LLM-Based Multi-Agent Framework for GitHub Issue Resolution,« *arXiv preprint arXiv:2403.17927,* 2024. |
| [81] | R. Loynd, »AutoGen's Teachable Agents,« AutoGen, 26 October 2023. [Mrežno]. Available: https://microsoft.github.io/autogen/blog/2023/10/26/TeachableAgent/. [Pokušaj pristupa 18 July 2024]. |
| [82] | Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger i C. Wang, »AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation,« 2023. |
| [83] | G. Trajkovski, »Behaviorism in Prompt Engineering: Shaping AI Responses,« 26 December 2023. [Mrežno]. Available: https://www.linkedin.com/pulse/behaviorism-prompt-engineering-shaping-ai-responses-goran-trajkovski-jyquc/. [Pokušaj pristupa 18 July 2024]. |
| [84] | K. Dumesnil, »The AI Fusionist Playbook for Learning Theories,« 6 October 2023. [Mrežno]. Available: https://www.linkedin.com/pulse/ai-fusionist-playbook-learning-theories-kambria-dumesnil/. [Pokušaj pristupa 18 July 2024]. |
| [85] | D. Moher, A. Liberati, J. Tetzlaff i D. Altman, »Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,« *INTERNATIONAL JOURNAL OF SURGERY,* svez. 8, br. 5, pp. 336-341, 2010. |
| [86] | N. Khan, Z. Khan, A. Koubaa, M. K. Khan i R. b. Salleh, »Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis,« *CONNECTION SCIENCE,* svez. 36, December 2024. |
| [87] | V. Salikutluk, D. Koert i F. Jaekel, »Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems,« u *HHAI 2023: AUGMENTING HUMAN INTELLECT*, 2023. |
| [88] | M. Lee, P. Liang i Q. Yang, »CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities,« u *PROCEEDINGS OF THE 2022 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI' 22)*, 2022. |
| [89] | J. Liu, »ChatGPT: perspectives from human-computer interaction and psychology,« *FRONTIERS IN ARTIFICIAL INTELLIGENCE,* svez. 7, June 2024. |
| [90] | Z. Zhou, J. Li, Z. Zhang, J. Yu i H. Duh, »Examining How the Large Language Models Impact the Conceptual Design with Human Designers: A Comparative Case Study,« *INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION,* 2024. |
| [91] | A. Casheekar, A. Lahiri, K. Rath, K. S. Prabhakar i K. Srinivasan, »A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions,« *COMPUTER SCIENCE REVIEW,* svez. 52, May 2024. |
| [92] | C. Qiu, Z. Xie, M. Liu i H. Hu, »Explainable Knowledge reasoning via thought chains for knowledge-based visual question answering,« *INFORMATION PROCESSING & MANAGEMENT,* svez. 61, July 2024. |
| [93] | L. Tang, P.-T. Jiang, H. Xiao i B. Li, »Towards Training-Free Open-World Segmentation via Image Prompt Foundation Models,« *INTERNATIONAL JOURNAL OF COMPUTER VISION,* 2024. |
| [94] | T. T. Procko, T. Elvira i O. Ochoa, »Dawn of the dialogue: AI's leap from lab to living room,« *FRONTIERS IN ARTIFICIAL INTELLIGENCE,* svez. 7, March 2024. |
| [95] | Y. Gao, W. Tong, E. Q. Q. Wu, W. Chen, G. Zhu i F.-Y. Wang, »Chat With ChatGPT on Interactive Engines for Intelligent Driving,« *IEEE TRANSACTIONS ON INTELLIGENT VEHICLES,* svez. 8, pp. 2034-2036, March 2023. |
| [96] | Y. Shu, H. Zhang, H. Gu, P. Zhang, T. Lu, D. Li i N. Gu, »RAH! RecSys-Assistant-Human: A Human-Centered Recommendation Framework With LLM Agents,« *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS,* 2024. |
| [97] | T. M. Al-Hasan, A. N. Sayed, F. Bensaali, Y. Himeur, I. Varlamis i G. Dimitrakopoulos, »From Traditional Recommender Systems to GPT-Based Chatbots: A Survey of Recent Developments and Future Directions,« *BIG DATA AND COGNITIVE COMPUTING,* svez. 8, April 2024. |
| [98] | J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, K. Zheng, D. Lian i E. Chen, »When large language models meet personalization: perspectives of challenges and opportunities,« *WORLD WIDE WEB-INTERNET AND WEB INFORMATION SYSTEMS,* svez. 27, July 2024. |
| [99] | M. J. Parker, C. Anderson, C. Stone i Y. Oh, »A Large Language Model Approach to Educational Survey Feedback Analysis,« *INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE IN EDUCATION,* 2024. |
| [100] | Y. Zhang, S. Gao, Y. Huang, Z. Yu i K. Tan, »3A-COT: an attend-arrange-abstract chain-of-thought for multi-document summarization,« *INTERNATIONAL JOURNAL OF MACHINE LEARNING AND CYBERNETICS,* 2024. |
| [101] | B. Gorer i P. B. Aydemir, »Generating Requirements Elicitation Interview Scripts with Large Language Models,« u *2023 IEEE 31ST INTERNATIONAL REQUIREMENTS ENGINEERING CONFERENCE WORKSHOPS, REW*, 2023. |
| [102] | N. Ngu, N. Lee i P. Shakarian, »Diversity Measures: Domain-Independent Proxies for Failure in Language Model Queries,« u *18TH IEEE INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING, ICSC 2024*, 2024. |
| [103] | T. Kojima, S. S. Gu, M. Reid, Y. Matsuo i Y. Iwasawa, »Large Language Models are Zero-Shot Reasoners,« u *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 35, NEURIPS 2022*, 2022. |
| [104] | Z. Shao, Z. Yu, M. Wang i J. Yu, »Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering,« u *2023 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)*, 2023. |
| [105] | T. Gao, A. Fisch i D. Chen, »Making Pre-trained Language Models Better Few-shot Learners,« u *59TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS AND THE 11TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (ACL-IJCNLP 2021), VOL 1*, 2021. |
| [106] | L. Reynolds i K. McDonell, »Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm,« u *EXTENDED ABSTRACTS OF THE 2021 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI'21)*, 2021. |
| [107] | T. Z. Zhao, E. Wallace, S. Feng, D. Klein i S. Singh, »Calibrate Before Use: Improving Few-Shot Performance of Language Models,« u *INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 139*, 2021. |
| [108] | Z. Quan i Z. Chen, »Human-computer pragmatics trialled: some (im)polite interactions with ChatGPT 4.0 and the ensuing implications,« *INTERACTIVE LEARNING ENVIRONMENTS,* 2024. |
| [109] | E. Y. Chang, »Prompting Large Language Models With the Socratic Method,« u *2023 IEEE 13TH ANNUAL COMPUTING AND COMMUNICATION WORKSHOP AND CONFERENCE, CCWC*, 2023. |
| [110] | M. Kochanek, I. Cichecki, O. Kaszyca, D. Szydlo, M. Madej, D. Jedrzejewski, P. Kazienko i J. Kocon, »Improving Training Dataset Balance with ChatGPT Prompt Engineering,« *ELECTRONICS,* svez. 13, June 2024. |
| [111] | H. Ma, C. Zhang, Y. Bian, L. Liu, Z. Zhang, P. Zhao, S. Zhang, H. Fu, Q. Hu i B. Wu, »Fairness-guided Few-shot Prompting for Large Language Models,« u *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 36 (NEURIPS 2023)*, 2023. |
| [112] | Y. Cheng, J. Chen, Q. Huang, Z. Xing, X. Xu i Q. Lu, »Prompt Sapper: A LLM-Empowered Production Tool for Building AI Chains,« *ACM TRANSACTIONS ON SOFTWARE ENGINEERING AND METHODOLOGY,* svez. 33, June 2024. |
| [113] | D. Park, G.-t. An, C. Kamyod i C. G. Kim, »A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model,« *JOURNAL OF WEB ENGINEERING,* svez. 22, pp. 1187-1206, 2023. |
| [114] | J. Seo, N. Zhang i C. Rong, »Flexible and Secure Code Deployment in Federated Learning using Large Language Models: Prompt Engineering to Enhance Malicious Code Detection,« u *2023 IEEE INTERNATIONAL CONFERENCE ON CLOUD COMPUTING TECHNOLOGY AND SCIENCE, CLOUDCOM 2023*, 2023. |
| [115] | T. Wu, E. Jiang, A. Donsbach, J. Gray, A. Molina, M. Terry i C. J. Cai, »PromptChainer: Chaining Large Language Model Prompts through Visual Programming,« u *EXTENDED ABSTRACTS OF THE 2022 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI 2022*, 2022. |
| [116] | B. Wang, G. Li i Y. Li, »Enabling Conversational Interaction with Mobile UI using Large Language Models,« u *PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI 2023*, 2023. |
| [117] | J. D. Zamfrescu-Pereira, R. Wong, B. Hartmann i Q. Yang, »Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts,« u *PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI 2023)*, 2023. |
| [118] | R. Bridgelall, »Unraveling the mysteries of AI chatbots,« *ARTIFICIAL INTELLIGENCE REVIEW,* svez. 57, March 2024. |
| [119] | C. Munley, A. Jarmusch i S. Chandrasekaran, »LLM4VV: Developing LLM-driven testsuite for compiler validation,« *FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE,* svez. 160, pp. 1-13, November 2024. |
| [120] | N. Marques, R. R. Silva i J. Bernardino, »Using ChatGPT in Software Requirements Engineering: A Comprehensive Review,« *FUTURE INTERNET,* svez. 16, June 2024. |
| [121] | S. H. Vemprala, R. Bonatti, A. Bucker i A. Kapoor, »ChatGPT for Robotics: Design Principles and Model Abilities,« *IEEE ACCESS,* svez. 12, pp. 55682-55696, 2024. |
| [122] | Y. Ye, H. You i J. Du, »Improved Trust in Human-Robot Collaboration With ChatGPT,« *IEEE ACCESS,* svez. 11, pp. 55748-55754, 2023. |
| [123] | H. Liu, Y. Zhu, K. Kato, A. Tsukahara, I. Kondo, T. Aoyama i Y. Hasegawa, »Enhancing the LLM-Based Robot Manipulation Through Human-Robot Collaboration,« *IEEE ROBOTICS AND AUTOMATION LETTERS,* svez. 9, pp. 6904-6911, August 2024. |
| [124] | X. Zhao, M. Li, C. Weber, M. B. Hafez i S. Wermter, »Chat with the Environment: Interactive Multimodal Perception Using Large Language Models,« u *2023 IEEE/RSJ INTERNATIONAL CONFERENCE ON INTELLIGENT ROBOTS AND SYSTEMS, IROS*, 2023. |
| [125] | D. T. T. Mai, C. V. Da i N. V. Hanh, »The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach,« *FRONTIERS IN EDUCATION,* svez. 9, February 2024. |
| [126] | I. C. Pelaez-Sanchez, D. Velarde-Camaqui i L. D. Glasserman-Morales, »The impact of large language models on higher education: exploring the connection between AI and Education 4.0,« *FRONTIERS IN EDUCATION,* svez. 9, June 2024. |
| [127] | G.-J. Hwang i N.-S. Chen, »Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions,« *EDUCATIONAL TECHNOLOGY & SOCIETY,* svez. 26, April 2023. |
| [128] | N. Kurian, »`No, Alexa, no!': designing child-safe AI and protecting children from the risks of the `empathy gap' in large language models,« *LEARNING MEDIA AND TECHNOLOGY,* 2024. |
| [129] | R. Zou, Z. Ye i C. Ye, »iTutor: A Generative Tutorial System for Teaching the Elders to Use Smartphone Applications,« u *ADJUNCT PROCEEDINGS OF THE 36TH ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE & TECHNOLOGY, UIST 2023 ADJUNCT*, 2023. |
| [130] | A. Bozkurt, »Unleashing the Potential of Generative AI, Conversational Agents and Chatbots in Educational Praxis: A Systematic Review and Bibliometric Analysis of GenAI in Education,« *OPEN PRAXIS,* svez. 15, pp. 261-270, 2023. |

# POPIS OZNAKA I KRATICA

|  |  |
| --- | --- |
| **AI** | Artificial Intelligence |
| **ANN** | Artificial Neural Network |
| **APE** | Automatic Prompt Engineer |
| **API** | application programming interface |
| **ART** | Automatic Reasoning and Tool-use |
| **CoT** | Chain-of-Thought |
| **DL** | Deep Learning |
| **GAN** | Generative adversarial networks |
| **GPT** | Generative Pre-trained Transformer |
| **HCAI** | Human-Centered AI |
| **HCI** | Human-Computer Interaction |
| **LLM** | Large Language Model |
| **ML** | Machine Learning |
| **NLP** | Natural Language Processing |
| **PAL** | Program-aided Language Models |
| **PRISMA** | Preferred Reporting Items for Systematic review and Meta-Analysis |
| **RAG** | Retrieval Augmented Generation |
| **ReAct** | Reasoning and Acting |
| **RLHF** | Reinforcement learning from Human Feedback |
| **TAM** | Technology Acceptance Model |
| **ToT** | Tree of Thoughts |
| **UI** | Umjetna inteligencija |
| **VAE** | Variational autoencoders |
| **WoS** | Web of Science |

# SAŽETAK

Ovaj rad istražuje interakciju sa sustavima umjetne inteligencije (UI) temeljenim na velikim jezičnim modelima. S razvojem velikih jezičnih modela, koji koriste napredne tehnike dubokog učenja i velike količine podataka, otvaraju se nove mogućnosti za interakciju između čovjeka i računala. Ključni fokus ovog istraživanja je na oblikovanjima upita za velike jezične modele kojima se optimiziraju odgovori modela, što omogućuje preciznije i korisnije rezultate.

Rad analizira osnovne postavke velikih jezičnih modela, različite tehnike oblikovanja upita te praktične primjere njihove primjene, uključujući usporedbe performansi nekoliko poznatih modela. Osim toga, rad se bavi širim aspektima interakcije sa sustavima UI, nudeći taksonomiju i smjernice za oblikovanje interakcije, te istražuje čovjeku usmjerenu UI, agentske sustave i bihevioristički pristup analizi sustava UI.

Kroz sustavan pregled literature, rad identificira interdisciplinarni karakter istraživanja u ovom području, koji obuhvaća računarstvo, psihologiju, obrazovanje, robotiku i inženjerstvo. Naglašava se potreba za daljnjim istraživanjem, osobito u pogledu razvoja naprednih tehnika oblikovanja upita, integracije velikih jezičnih modela s drugim tehnologijama UI te unapređenja transparentnosti i objašnjivosti odluka koje donose ovi modeli.

Zaključno, rad postavlja temelje za buduće istraživanje i razvoj u području interakcije sa sustavima UI temeljenim na velikim jezičnim modelima, s ciljem unapređenja načina na koji komuniciramo s tehnologijom i utjecaja koji ovi sustavi imaju na društvo u cjelini.

1. OpenAI https://openai.com/about/ [↑](#footnote-ref-2)
2. Anthropic https://www.anthropic.com/company [↑](#footnote-ref-3)
3. Meta https://about.meta.com/ [↑](#footnote-ref-4)