

# Inženjering upita za velike jezične modele

---

**Martinović, Ivan**

**Master's thesis / Diplomski rad**

**2024**

*Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj:* **Josip Juraj Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek / Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek**

*Permanent link / Trajna poveznica:* <https://um.nsk.hr/um:nbn:hr:200:217763>

*Rights / Prava:* [In copyright](#) / [Zaštićeno autorskim pravom.](#)

*Download date / Datum preuzimanja:* **2024-11-27**

*Repository / Repozitorij:*

[Faculty of Electrical Engineering, Computer Science and Information Technology Osijek](#)



**SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU  
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I  
INFORMACIJSKIH TEHNOLOGIJA OSIJEK**

**Sveučilišni studij**

**INŽENJERING UPITA ZA VELIKE JEZIČNE MODELE**

**Diplomski rad**

**Ivan Martinović**

**Osijek, 2024.**

# SADRŽAJ

<b>1. UVOD</b> .....	<b>1</b>
<b>2. PREGLED PODRUČJA</b> .....	<b>3</b>
<b>2.1. Općenito o obradi prirodnog jezika</b> .....	<b>3</b>
<b>2.2. Definicija i svrha obrade prirodnog jezika</b> .....	<b>3</b>
<b>2.3. Povijest i razvoj područja</b> .....	<b>3</b>
<b>2.4. Znanstveni radovi koji su utjecali na razvoj velikih jezičnih modela</b> .....	<b>5</b>
2.4.1. "Attention is All You Need" .....	5
2.4.2. BERT ( <i>engl. Bidirectional Encoder Representations from Transformers</i> ) .....	5
2.4.3. GPT ( <i>engl. Generative Pre-Trained Transformers</i> ) .....	6
<b>2.5. Veliki jezični modeli (eng. <i>Large language models (LLM)</i>)</b> .....	<b>6</b>
<b>2.6. Predtreniranje jezičnih modela</b> .....	<b>9</b>
2.6.1. Predtreniranje korpus teksta .....	9
2.6.2. Maskirano jezično modeliranje ( <i>engl. Masked language modelling (MLM)</i> ).....	9
2.6.3. Predviđanje sljedeće rečenice ( <i>engl. Next sentence prediction (NSP)</i> ).....	11
2.6.4. Predviđanje sljedećeg tokena ( <i>engl. Next token prediction (NTP)</i> ) .....	12
2.6.5. Samonadzirano učenje ( <i>engl. Self-supervised learning (SSL)</i> ) .....	12
<b>2.7. Razlike između BERT i GTP modela</b> .....	<b>13</b>
2.7.1. GPT metodologija treniranja .....	13
2.7.2. BERT metodologija treniranja.....	14
<b>3. TRANSFORMERI</b> .....	<b>16</b>
<b>3.1. Uvod u transformere</b> .....	<b>16</b>
<b>3.2. Arhitektura transformera</b> .....	<b>16</b>
<b>3.3. Primjena transformera u obradi prirodnog jezika</b> .....	<b>28</b>
<b>3.4. Prednosti i ograničenja transformera</b> .....	<b>30</b>
<b>4. MEHANIZAM PAŽNJE U STROJNOM UČENJU (<i>ENGL. ATTENTION MECHANISM IN MACHINE LEARNING</i>)</b> .....	<b>32</b>
<b>4.1. Vrste mehanizama pažnje</b> .....	<b>32</b>
<b>4.2. Vrste prozora pažnje (<i>engl. Attention window</i>)</b> .....	<b>33</b>
<b>4.3. Razlike u treniranju bidirekcionalnih i dekodek-samo modela (BERT vs. GPT)</b> .....	<b>34</b>
<b>4.4. Jezični modeli i njihove razlike</b> .....	<b>35</b>

4.5. Skupovi podataka za treniranje.....	36
4.6. Zaključak.....	38
<b>5. INŽENJERING UPITA (<i>engl. Prompt Engineering</i>).....</b>	<b>39</b>
5.1. Opće postavke velikih jezičnih modela .....	39
5.2. Inženjering upita s nula pokušaja .....	40
5.3. Inženjering upita s više pokušaja .....	45
5.4. Inženjering upita s lančanim promišljanjem.....	50
5.5. Generiranje uz dohvaćanje podataka ( <i>engl. Retrieval augmented generation (RAG)</i> ).....	54
5.6. Inženjering upita s poticanjem zaključivanja ( <i>engl. ReAct Prompting</i> ).....	58
5.7. Koncept i svrha inženjeringa upita .....	61
<b>6. ZAKLJUČAK.....</b>	<b>62</b>
<b>LITERATURA .....</b>	<b>63</b>
<b>SAŽETAK.....</b>	<b>68</b>
<b>ABSTRACT .....</b>	<b>69</b>
<b>ŽIVOTOPIS.....</b>	<b>70</b>
<b>PRILOZI.....</b>	<b>71</b>

## 1. UVOD

U području obrade prirodnog jezika i umjetne inteligencije, razvoj naprednih modela obrade prirodnog jezika donio je eksponencijalni napredak u razumijevanju i stvaranju prirodnog jezika. Među tim modelima najistaknutiji su transformeri [1], koji su postali ključna tehnologija u mnogim zadacima obrade prirodnog jezika, kao što je generiranje teksta, odgovaranje na korisničke upite, strojno prevođenje itd. Međutim, iako su transformeri vrlo sofisticirani i korisni, suočavaju se s izazovima pri kontroli generiranih izlaza.

Upravo tu dolazi do izražaja koncept inženjeringa upita [2]. Inženjering upita je pristup koji se koristi za usmjeravanje i kontrolu generiranja teksta jezičnih modela. Ovakav pristup omogućuje korisnicima precizniju specifikaciju očekivanih izlaznih modela i željene rezultate. Važnost inženjeringa upita proizlazi iz činjenice gdje je generiranje teksta temeljna komponenta mnogih aplikacija s umjetnom inteligencijom, uključujući chatbotove, automatizirane sustave odgovaranja na upite, generiranje sažetaka ili čak pisanje članaka. Upravljanje generiranim rezultatom postaje ključno za osiguranje preciznosti, dosljednosti i usklađenosti sa željenim kriterijima.

Pristupi inženjeringa upita uključuju dizajniranje i prilagodbu upita (uputa) koje se koriste kao ulazni signali za jezične modele. Upute mogu biti kratke fraze, pitanja ili neki drugi oblik specifičnih uputa. Eksperimentiranjem s različitim uputama moguće je kontrolirati izlazni slijed modela, poboljšati razumljivost ili fokusirati na određenu temu.

S obzirom na složenost i raznolikost prirodnog jezika, inženjering upita postaje nužan alat za precizno oblikovanje upita koji utječu na kvalitetu izlaznih podataka. Ova praksa također potiče kreativnu upotrebu jezičnih modela, omogućujući korisnicima istraživanje različite načine interakcije s AI tehnologijama. S aspekta dizajna upita, važno je razumjeti kako određeni izrazi i konstrukcije utječu na interpretaciju i odgovor modela, što zahtijeva detaljnu analizu i razumijevanje kako jezični modeli procesuiraju jezik.

Diplomski rad za cilj ima dati pregled inženjeringa upita, istražiti njegovu izvedivost i objasniti metode i tehnike korištene u njegovoj primjeni. Također, razmotrit će se najnovije AI tehnologije, kao što su GPT-3/4 [3], Llama [4], Chinchilla [5] i Gemini [6], te analizirati njihovu primjenu i rezultate u kontekstu inženjeringa upita.

Nadalje, rad će istražiti kako inženjering upita utječe na smanjenje pristranosti i povećanje objektivnosti u generiranom tekstu. Budući da jezični modeli uče iz velikih baza podataka,

postoji rizik od učenja i ponavljanja pristranih obrazaca. Precizno oblikovani upiti mogu pomoći u minimiziranju ovih problema, usmjeravajući modele prema neutralnijem i objektivnijem generiranju teksta.

Uz to, ovaj rad će analizirati utjecaj inženjeringa upita na poboljšanje interakcije između čovjeka i stroja, posebno u kontekstu korisničkog iskustva. Kako se AI tehnologije sve više integriraju u svakodnevni život, sve važnija je sposobnost jezičnih modela da učinkovito komuniciraju i odgovaraju na korisničke zahtjeve. Kroz analizu primjera iz prakse, ovaj rad će prikazati kako prilagođeni upiti mogu poboljšati razumijevanje i zadovoljstvo korisnika u interakcijama s AI sustavima.

## **2. PREGLED PODRUČJA**

### **2.1. Općenito o obradi prirodnog jezika**

Obrada prirodnog jezika [7] je pod područje računarske znanosti, obrade informacija i umjetne inteligencije koja proučava interakciju između računala i ljudskih (prirodnih) jezika, tj. jezikom kojim se ljudi svakodnevno koriste za komunikaciju. Glavna svrha obrade prirodnog jezika je omogućiti računalima da razumiju, generiraju i manipuliraju jezikom na način sličan ljudskom, tj. raspoznavanje govora, razumijevanje prirodnog jezika i generiranje prirodnog jezika.

### **2.2. Definicija i svrha obrade prirodnog jezika**

Obrada prirodnog jezika (engl. Natural Language Processing – NLP [8]) je interdisciplinarno područje računalne znanosti, umjetne inteligencije i lingvistike koje se bavi obradom prirodnog jezika, tj. oblikom komunikacije koje ljudi koriste za izražavanje misli i ideja. Definiira se kao skup tehnika, metoda i algoritama koji omogućuju računalima da razumiju, tumače i generiraju jezične strukture slično kao ljudi.

Primjeri primjene obrade prirodnog jezika su mnogobrojni. Filtriranje nepoželjnih poruka u elektroničkoj pošti [9] temelji se na analizi jezičnih obrazaca i klasifikaciji poruka kao legitimne ili neželjene. Strojno prevođenje je još jedan primjer jezične obrade, gdje se tekst prevodi s jednog jezika na drugi. Automatsko generiranje sažetaka tekstova, prepoznavanje govora, analiza konteksta na društvenim mrežama, chatbotovi, pretraživanje informacija - sve su to primjeri primjene obrade prirodnog jezika.

### **2.3. Povijest i razvoj područja**

Obrada prirodnog jezika potječe iz ranog razdoblja računalne znanosti, posebno iz područja umjetne inteligencije. Pioniri poput Alana Turinga [10] i Noama Chomskog [11] razmatrali su koncepte vezane uz obradu prirodnog jezika već u 20. stoljeću. No, tek s razvojem moćnijih računala i dostupnosti velikih skupova jezičnih podataka, obrada prirodnog jezika postala je značajna i sveprisutna domena istraživanja.

Ovo je samo nekoliko primjera ranih prevoditelja temeljenih na obradi prirodnog jezika koji su otvorili put za napredak u strojnom prevođenju koji vidimo danas.

Razvoj obrade prirodnog jezika postupno je praćen napretkom računalne snage i dostupnosti velikih jezičnih skupova podataka. Rani sustavi za strojno prevođenje pojavili su se već u 1950-

ima, ali su se s vremenom poboljšavali i prilagođavali novim tehnikama i algoritmima. U posljednjem desetljeću su se pojavili revolucionarni jezični modeli poput GPT-3 (Generative Pre-trained Transformer 3) [2] i BERT (Bidirectional Encoder Representations from Transformers) [12] koji su postigli izvanredne rezultate u razumijevanju i generiranju jezičnih struktura.

Obrada prirodnog jezika razvijala se paralelno s napretkom računalnih tehnologija i sve većom dostupnošću velikih jezičnih skupova podataka. U ranim fazama razvoja obrade prirodnog jezika, naglasak je bio na pravilnom prepoznavanju i razumijevanju pojedinih riječi i gramatičkih struktura. No, s vremenom su se tehnike proširile i obuhvatile složenije aspekte jezika poput semantike, pragmatike i konteksta.

Jedan od primjera razvoja obrade prirodnog jezika je računalni prevoditelj SYSTRAN [13], koji je razvijen sredinom 1970-ih. SYSTRAN je koristio rječnik i gramatička pravila za prevođenje teksta s jednog jezika na drugi. U kasnijim godinama, sa sve većim količinama dostupnih paralelnih tekstova, primjena statističkih metoda i strojnog učenja, postala je dominantna u području strojnog prevođenja.

Razvoj obrade prirodnog jezika doživio je značajan proboj s dolaskom jezičnih modela temeljenih na transformer arhitekturi. GPT (Generative Pre-trained Transformer) [2] modeli, poput GPT-3 koji je predstavljen 2020. godine, postigli su izvanredne rezultate u generiranju jezičnih struktura. Ovi modeli se prethodno treniraju na velikim skupovima podataka i zatim se prilagođavaju za specifične zadatke kao što su strojno prevođenje, generiranje teksta ili odgovaranje na upite.

Uz sve brži napredak tehnologije i dostupnost velikih količina jezičnih podataka, obrada prirodnog jezika nastavlja napredovati. Napredne tehnike strojnog učenja, dubokog učenja i obrade prirodnog jezika omogućuju računalima sve složenije razumijevanje jezika i generiranje relevantnih i kontekstualno bogatih jezičnih struktura.

Obrada prirodnog jezika neprekidno napreduje i nastavlja pružati temelj za brojne inovativne aplikacije u području računalne znanosti i umjetne inteligencije. Sve veća sposobnost računala da razumiju, tumače i generiraju jezične strukture približava nas ostvarenju cilja da računala komuniciraju s ljudima na prirodan način, otvarajući put za razvoj naprednih sustava za prevođenje, automatizirane asistente, analizu konteksta, automatsko generiranje teksta i još puno toga.



Napredak u obradi prirodnog jezika pruža nove mogućnosti za poboljšanje svakodnevne interakcije s računalima i pruža osnovu za razvoj novih tehnologija koje će transformirati način na koji koristimo jezik u digitalnom okruženju.

## **2.4. Znanstveni radovi koji su utjecali na razvoj velikih jezičnih modela**

### **2.4.1. "Attention is All You Need"**

Rad "Attention is All You Need" [1] autora Vaswanija i suradnika (2017.) revolucionirao je obradu prirodnog jezika uvođenjem transformer modela, koji se u potpunosti oslanja na mehanizme pozornosti, izbacujući slojeve rekurentne i konvolucijske mreže korištene u prethodnim modelima. Transformer koristi mehanizam samo-pažnje za obradu svih tokena u nizu istovremeno, omogućujući paralelizaciju i značajno poboljšanje učinkovitosti i performansi u zadacima strojnog prevođenja. Arhitektura ovog modela postala je temelj za mnoge napredne jezične modele, uključujući BERT i GPT, zbog svoje sposobnosti učinkovite obrade dugoročnih ovisnosti u tekstu.

Ključna inovacija transformer modela je upotreba mehanizama višestruke samo-pažnje, što omogućuje modelu da se istovremeno fokusira na različite dijelove rečenice, razumijevajući razne jezične značajke. Osim toga, model koristi pozicijske ugradnje za održavanje reda tokena, kompenzirajući nedostatak povratne veze.

### **2.4.2. BERT (*engl. Bidirectional Encoder Representations from Transformers*)**

BERT [12] predstavljen od strane Devlina i suradnika (2019.), predstavlja značajan iskorak u obradi prirodnog jezika primjenom dvosmjernog treniranja na transformer modelu. Za razliku od modela koji obrađuju tekst u jednom smjeru (ili s lijeva na desno ili s desna na lijevo), BERT istovremeno uzima u obzir oba smjera tijekom treniranja, omogućujući mu da razumije puni kontekst riječi na temelju njezinih susjednih riječi. Ovaj dvosmjerni pristup učinio je BERT posebno učinkovitim u zadacima poput odgovaranja na pitanja, analize konteksta i prepoznavanja entiteta.

BERT-ovo treniranje uključuje dva ključna koraka: maskirano učenje jezika (*engl. masked language modeling (MLM)*) i predviđanje sljedeće rečenice (*engl. next sentence prediction (NSP)*). U MLM-u, nasumične riječi u rečenici su maskirane, a model se trenira kako bi ih predvidio, što ga prisiljava da nauči kontekst riječi iz oba smjera. U NSP-u, model uči logički predvidjeti slijedi li određena rečenica nakon druge, poboljšavajući svoje razumijevanje odnosa

između rečenica. Inovativan pristup BERT-a doveo je do vrhunskih performansi čineći ga ključnim elementom u razvoju suvremenih NLP modela.

### **2.4.3. GPT (engl. *Generative Pre-Trained Transformers*)**

GPT [14] razvijen od strane Radforda i suradnika (2018., 2019.), predstavlja revolucionarni model u obradi prirodnog jezika koji kombinira nenadzirano prethodno treniranje s nadzirim finim podešavanjem. Arhitektura GPT-a temelji se na transformer modelu, ali se razlikuje od BERT-a po tome što je jednosmjernan, obrađujući tekst s lijeva na desno. Tijekom predtreniranja, GPT uči jezične obrasce iz velikih količina teksta na nenadzirani način, omogućujući mu generiranje koherentnog i kontekstualno relevantnog teksta. Ova sposobnost čini GPT posebno pogodnim za zadatke poput dovršavanja teksta, generiranja dijaloga i kreativnog pisanja.

Trening GPT-a odvija se u dvije faze: prvo, model prolazi kroz nenadzirano predtreniranje na velikom tekstualnom korpusu, gdje uči predviđati sljedeću riječ u nizu. Nakon toga slijedi nadzirano fino podešavanje na manjem, specifičnom skupu podataka kako bi se model prilagodio specifičnim aplikacijama kao što su prevođenje, sažimanje ili odgovaranje na pitanja. Uspjeh GPT-a u generiranju teksta nalik ljudskom doveo je do njegove primjene u raznim područjima, a njegove kasnije verzije, poput GPT-3, GPT-3T, i GPT-4o, nastavile su pomicati granice onoga što je moguće postići jezičnim modelima.

## **2.5. Veliki jezični modeli (eng. *Large language models (LLM)*)**

Veliki jezični modeli (eng. *Large Language Models*) [15] napredni su algoritmi umjetne inteligencije koji se obučavaju na velikim količinama tekstualnih podataka kako bi stekli sposobnost razumijevanja i generiranja prirodnog jezika. Ti modeli koriste neuronske mreže s milijardama parametara za obradu i tumačenje složenih jezičnih obrazaca. Glavna karakteristika velikih jezičnih modela je njihova sposobnost obavljanja različitih jezičnih zadataka, uključujući prevođenje, generiranje teksta, odgovaranje na pitanja i još mnogo toga.

Ove napredne tehnologije temelje se na različitim karakteristikama:

- **Obrada konteksta i semantike:** Veliki jezični modeli sposobni su razumjeti kontekst i značenje rečenica, što im omogućuje generiranje koherentnog i smislenog teksta. Na primjer, ako se postavi upit: "Koja je sada godina?", veliki jezični model će uzeti u obzir kontekst i dati točan odgovor.
- **Generiranje tekstualnog sadržaja:** Veliki jezični modeli mogu generirati tekst visoke kvalitete koji je sličan ljudskom. Na primjer, mogu stvarati eseje, priče, članke i druge oblike pisanog

materijala. To otvara mogućnosti za automatizaciju pisanja ili generiranje personaliziranog sadržaja.

- Prilagodba različitim zadacima: Veliki jezični modeli su fleksibilni i mogu se prilagoditi različitim zadacima za obradu prirodnog jezika. Bilo da se radi o prevodenju ili odgovaranju na pitanja, modeli mogu naučiti određene obrasce i pružiti relevantne rezultate.

Primjeri poznatih velikih jezičnih modela

- Jedan od najpoznatijih primjera velikih jezičnih modela (eng. *large language models*) je GPT-3 (eng. *Generative Pre-trained Transformer 3*) [2], [16], koji je razvio istraživački laboratorij OpenAI. GPT-3 je iznimno velik model sa više od 175 milijardi parametara, što mu daje iznimne mogućnosti jezične obrade. Može generirati tekstualni sadržaj, odgovarati na pitanja, prevoditi jezik i obavljati druge složene zadatke. Primjeri najnovijih GPT modela su GPT-4 [17] i GPT-4o [18] koji su multi-modalni modeli, što znači kako mogu prihvatiti i tekstualni i grafički ulaz dok je izlaz tekstualni.
- BERT (eng. *Bidirectional Encoder Representations from Transformers*) [12]: BERT je popularan veliki jezični model koji je razvio Google. Ovaj se model temelji na transformerskoj arhitekturi i treniran je na velikim setovima tekstualnih podataka. BERT ima sposobnost razumijevanja konteksta i semantike teksta i koristi se za različite lingvističke zadatke kao što su pretraživanje, prepoznavanje entiteta, klasifikacija i drugo.
- RoBERTa (robustno optimizirani BERT) [19]: RoBERTa je varijanta BERT-a koja se temelji na sličnoj arhitekturi, ali dodatno optimizirana za bolju izvedbu u raznim zadacima obrade prirodnog jezika. RoBERTa koristi napredne tehnike prekvalifikacije za postizanje boljih rezultata u razumijevanju i generiranju teksta.
- DistilBERT [20] je mali, brz, jeftin i lagan transformer model temeljen na BERT arhitekturi. Tijekom faze predtreniranja provodi se destilacija znanja kako bi se veličina BERT modela smanjila i do 40%. Kako bi se iskoristile induktivne pristranosti koje su veći modeli naučili tijekom predtreniranja, autori uvode trostruki gubitak koji kombinira jezično modeliranje, destilaciju i gubitke kosinusne udaljenosti.
- ALBERT [21] je kompaktan, učinkovit i prilagodljiv jezični model temeljen na BERT arhitekturi. Optimizacija parametara provodi se tijekom faze predtreniranja kako bi se smanjila veličina BERT modela i poboljšala njegova brzina obrade. ALBERT postiže visoku razinu performansi primjenom mehanizama dijeljenja parametara i faktorizacije

matrice učenja, čime se smanjuje broj parametara bez gubitka preciznosti. Autori također uvode novu metodu regularizacije kako bi dodatno unaprijedili učenje modela.

- T5 (*engl. Text-to-Text Transfer Transformer*) [22]: T5 je veliki jezični model koji je razvio Google koji se ističe svojom sposobnošću prijenosa različitih jezičnih zadataka na jedinstven način. Umjesto da se modelira za svaki zadatak zasebno, T5 je osposobljen za prevođenje jednog oblika teksta u drugi, što mu omogućuje primjenu na različite zadatke kao što su prevođenje, generiranje teksta i odgovaranje na pitanja.

Primjena velikih jezičnih modela moguća je u različitim područjima, uključujući:

- Medijska industrija: U medijskoj industriji veliki jezični modeli koriste se za automatizaciju generiranja članaka, koji mogu uključivati sportske vijesti, financijska izvješća ili čak tehnološke vijesti. Također se koriste za izradu personaliziranih preporuka sadržaja prilagođenih interesima korisnika. Na primjer, algoritmi temeljeni na velikim jezičnim modelima mogu prepoznati preferencije gledatelja na temelju pregledanog sadržaja i preporučiti im sličan sadržaj.
- Zdravstvo: Veliki jezični modeli imaju potencijal transformirati zdravstvenu industriju. Oni se mogu koristiti za analizu medicinskih podataka kako bi se lakše pronašli obrasci ili identificirali potencijalni rizici od bolesti. Oni također mogu pružiti informacije o dijagnozama i terapijama. Na primjer, modeli poput GPT-4o mogu se obučiti na medicinskoj literaturi i istraživanjima te dati stručne preporuke na temelju najnovijih saznanja.
- Obrazovanje: Veliki jezični modeli mogu pomoći učenicima i nastavnicima u obrazovanju. Na primjer, mogu se koristiti za automatsko ocjenjivanje zadataka, pružajući brže povratne informacije studentima. Oni također mogu generirati obrazovne materijale i odgovarati na pitanja učenika, pružajući dodatnu podršku učenju.
- Pravosuđe [23]: Primjena obrade prirodnog jezika u pravosuđu revolucionira način na koji pravnici pristupaju analizi i istraživanju pravnih tekstova. Alati za obradu prirodnog jezika omogućuju automatsko pretraživanje i kategorizaciju pravnih dokumenata, identifikaciju ključnih pravnih pojmova i ekstrakciju relevantnih informacija iz velikih količina podataka. Također, obrada prirodnog jezika koristi se i za analizu presuda, izradu sažetaka i predviđanje pravnih ishoda na temelju prethodnih slučajeva. Ova tehnologija poboljšava efikasnost pravne prakse i omogućuje pravnicima da donose bolje informirane odluke uz uštedu vremena i resursa.

Gore navedeni primjeri samo su neki od mnogih načina na koje se veliki jezični modeli mogu primijeniti u stvarnom svijetu. Njihov potencijal za inovacije i transformaciju raznih industrija i sektora još uvijek se istražuje i razvija.

## **2.6. Predtreniranje jezičnih modela**

Predtreniranje jezičnih modela odnosi se na početnu fazu treniranja jezičnog modela na velikom skupu podataka kako bi model naučio osnovne jezične obrasce i strukture. Ova faza je ključna jer omogućava modelu da stvori bogatu unutarnju reprezentaciju jezika prije nego što se prilagodi specifičnim zadacima.

### **2.6.1. Predtreniranje korpus teksta**

Poznato je da poboljšanje kvalitete podataka prije predtreniranja jezičnih modela, ili odabir podataka iz domena koje su slične zadacima za kasniju obradu, obično vodi ka boljoj izvedbi modela. Međutim, uloga koju kvaliteta korpusa, posebno njegova kompleksnost, ima na kasniju izvedbu modela još uvijek nije dovoljno istražena. Opsežne empirijske analize [24], [25], [26] koju su provedene pokazale su da takav raznolik i kompleksan korpus značajno poboljšava izvedbu modela u usporedbi s početnim korpusima manje raznolikosti i kompleksnosti, posebno u kontekstu zadataka općeg jezičnog razumijevanja.

### **2.6.2. Maskirano jezično modeliranje (*engl. Masked language modelling (MLM)*)**

Maskirano jezično modeliranje [27], [28] (*engl. Masked language modelling (MLM)*) je tehnika koja se koristi tijekom predtreniranja gdje se određeni postotak riječi u rečenici maskira (zamjenjuje posebnim tokenom) [29] i zadatak modela je da predvidi te maskirane riječi na temelju konteksta.

Model se trenira na velikim količinama tekstualnih podataka kako bi naučio razumijevanje konteksta riječi i predviđanje maskiranih tokena na temelju njihovog okruženja.

Na primjer, za rečenicu "Pas [MASK] kost", model bi predvidio riječ "jede" kao maskirani token. Tijekom procesa treniranja, model se ažurira na temelju razlike između svojih predviđanja i stvarnih riječi u rečenici. Ova faza predtreniranja pomaže modelu naučiti korisne kontekstualne reprezentacije riječi, koje se zatim mogu fino podesiti za specifične zadatke obrade prirodnog jezika. Cilj MLM-a je iskoristiti velike količine dostupnih tekstualnih podataka za učenje modela jezika opće namjene koji se mogu primijeniti na različite probleme u obradi prirodnog jezika.

## Primjena maskiranog jezičnog modeliranja

Modeliranje maskiranog jezika se etabliralo kao moćna tehnika s širokim rasponom primjena u domeni obrade prirodnog jezika. Njegova sposobnost učenja iz neoznačenih podataka i pronalaženja dugoročnih ovisnosti u jeziku čini ga vrijednim alatom za razne zadatke u obradi prirodnog jezika.

### 1. Odgovaranje na pitanja

Modeliranje maskiranog jezika može se koristiti za predtreniranje modela za zadatke odgovaranja na pitanja. Model uči identificirati relevantan odgovor na postavljeno pitanje unutar danog konteksta.

### 2. Prepoznavanje imenovanih entiteta

Modeliranje maskiranog jezika može se koristiti za predtreniranje modela za zadatke prepoznavanja imenovanih entiteta. Model uči identificirati i kategorizirati imenovane entitete u tekstu, kao što su imena ljudi, organizacije i geografske lokacije. Modeli poput BERT-a i RoBERTa su postigli vrhunske rezultate na NER [30], [31] benchmarkovima poput CoNLL-2003.

### 3. Generiranje teksta

Modeliranje maskiranog jezika može se koristiti za predtreniranje modela za zadatke generiranja teksta. Model uči generirati koherentan i smislen tekst na temelju danog upita ili početnog teksta. Modeli poput GPT-3 i GPT-4o su postigli impresivne rezultate u generiranju različitih kreativnih tekstualnih formata, poput pjesama, skripti, članaka i slično.

### 4. Strojno prevođenje

Modeliranje maskiranog jezika može se koristiti za predtreniranje modela za zadatke strojnog prevođenja. Model uči prevoditi tekst s jednog jezika na drugi, očuvanjem originalnog značenja i konteksta.

### 5. Mijenjanje članaka

Mijenjanje članaka je tehnika koja uključuje stvaranje novih varijacija postojećih članaka mijenjanjem riječi, fraza ili rečenica kako bi se generirao novi sadržaj sličnog značenja kao original. Ova se tehnika često koristi u SEO svrhe, s ciljem stvaranja više verzija

članka koje mogu ciljati različite ključne riječi ili fraze. Iako se ova tehnika često koristi u komercijalne svrhe, može se primijeniti i u znanstvenom kontekstu za generiranje različitih varijacija teksta za analizu ili usporedbu.

Modeliranje maskiranog jezika pokazao se kao sveobuhvatan pristup koji poboljšava performanse modela u raznim zadacima obrade prirodnog jezika. Njegova sposobnost učenja iz velikih količina neoznačenih podataka i pronalaženja dugoročnih ovisnosti u jeziku čini ga vrijednim alatom u domeni obrade prirodnog jezika.

### **2.6.3. Predviđanje sljedeće rečenice (*engl. Next sentence prediction (NSP)*)**

Predviđanje sljedeće rečenice [32] (Next Sentence Prediction - NSP) je ključna tehnika korištena u treniranju jezičnih modela, posebno unutar arhitektura kao što su BERT. Ova tehnika pomaže modelima da razumiju koherentnost i logički slijed rečenica unutar teksta, što je od velikog značaja za mnoge zadatke obrade prirodnog jezika (NLP). BERT model je treniran upravo na ovom zadatku predviđanja sljedeće rečenice (NSP) u kombinaciji s maskiranjem riječi (Masked Language Modeling - MLM).

NSP zadatak uključuje predočavanje parova rečenica gdje model mora odrediti jesu li te dvije rečenice susjedne u izvornom tekstu ili su nasumično sparene. Prva rečenica se označava kao "rečenica A", a druga kao "rečenica B". Model se trenira na način da predvidi vjerojatnost da rečenica B slijedi rečenicu A.

Modeli se treniraju koristeći velike korpuse tekstova gdje su rečenice nasumično sparene kako bi se kreirali negativni primjeri [33]. Pozitivni primjeri su uzeti direktno iz izvornog teksta. Tijekom treniranja, model optimizira svoje predikcije uspoređujući ih s stvarnim oznakama koje indiciraju jesu li rečenice susjedne ili ne. Cilj je minimizirati gubitak, koji predstavlja razliku između predikcija modela i stvarnih oznaka.

Predviđanje sljedeće rečenice (NSP) je esencijalna komponenta treniranja jezičnih modela koja pomaže u razumijevanju logičkog slijeda i koherencije teksta. Kroz korištenje velikih korpusa tekstova i optimizaciju na temelju stvarnih oznaka, NSP omogućava modelima da razviju dublje razumijevanje konteksta, čime se poboljšavaju njihove performanse u različitim NLP zadacima. Daljnja istraživanja mogla bi se usmjeriti na optimizaciju NSP zadatka kako bi se dodatno poboljšala učinkovitost modela u specifičnim primjenama.

#### **2.6.4. Predviđanje sljedećeg tokena (*engl. Next token prediction (NTP)*)**

Predviđanje sljedećeg tokena (Next Token Prediction - NTP) [34] je temeljna tehnika u treniranju jezičnih modela, posebno u generativnim modelima poput GPT-3 (Generative Pre-trained Transformer 3). Ova tehnika omogućava modelima da generiraju tekst predviđanjem sljedećeg tokena u sekvenci na temelju prethodnih tokena.

NTP zadatak uključuje treniranje modela da predvidi sljedeći token (riječ, interpunkcijski znak ili drugi simbol) u nizu tokena na temelju konteksta prethodnih tokena. Model se iterativno trenira na velikim skupovima podataka kako bi naučio obrasce i strukture jezika, omogućujući mu da generira koherentan i smislen tekst.

Tijekom treninga, model se hrani velikim korpusima tekstualnih podataka. Svaka rečenica se razbija u tokene, a model se trenira da predvidi svaki sljedeći token u sekvenci. Gubitak se računa kao razlika između predviđenih tokena i stvarnih tokena, a model optimizira svoje parametre kako bi minimizirao ovaj gubitak.

NTP je ključan za generativne zadatke u obradbi prirodnog jezika, kao što su generiranje teksta, dovršavanje rečenica i automatski odgovori na pitanja. Modeli trenirani na NTP zadatku mogu generirati tekst koji je koherentan i prirodan, simulirajući ljudski jezik. Ovo omogućava njihovu primjenu u širokom rasponu aplikacija, uključujući chatbotove, asistente za pisanje i automatsko generiranje sadržaja.

#### **2.6.5. Samonadzirano učenje (*engl. Self-supervised learning (SSL)*)**

Samonadzirano učenje [35] je inovativna metoda treniranja modela posebno popularna u području obrade prirodnog jezika (NLP). Ova metoda omogućava modelima da uče iz neoznačenih podataka, koristeći same podatke kao nadzor.

U samonadziranom učenju, model se trenira na način da rješava zadatke koji se generiraju iz neoznačenih podataka. To znači da model koristi informacije sadržane u ulaznim podacima za stvaranje nadzornih signala. Primjeri zadataka samonadziranog učenja uključuju maskirano jezično modeliranje (MLM), predviđanje sljedeće rečenice (NSP) i predviđanje sljedećeg tokena (NTP).

Modeli se treniraju na velikim korpusima neoznačenih tekstualnih podataka. Svaki korpus se koristi za generiranje nadzornih signala bez potrebe za ručnim označavanjem. Gubitak se računa kao razlika između stvarnih vrijednosti (*engl. ground truth*) i predviđenih vrijednosti, a model optimizira svoje parametre kako bi minimizirao ovaj gubitak.



Samonadzirano učenje omogućava treniranje modela na velikim količinama podataka bez potrebe za ručnim označavanjem, što značajno smanjuje troškove i vrijeme potrebno za pripremu podataka. Ova metoda također omogućava modelima da uče bogate i složene reprezentacije jezika, što poboljšava njihovu sposobnost generalizacije i poboljšava performanse na različitim zadacima obrade prirodnog jezika, uključujući prepoznavanje imenovanih entiteta, generiranje teksta i strojno prevođenje.

Razumijevanje ovih ključnih pojmova i tehnika je ključno za dublje razumijevanje kako moderni jezični modeli postižu visoku učinkovitost i prilagodljivost. Predtreniranje na velikim korpusima teksta koristeći metode poput MLM-a i NSP-a, u kombinaciji sa NTP-om, omogućava stvaranje moćnih alata za jezičnu obradu i analizu.

## 2.7. Razlike između BERT i GTP modela

### 2.7.1. GPT metodologija treniranja

GPT model slijedi dvostupanjsku metodologiju treninga, naglašavajući neovisno učenje, a zatim nadzirano fino podešavanje [14]. Ova metodologija osmišljena je kako bi se maksimizirala sposobnost modela za generiranje koherentnog i kontekstualno prikladnog teksta.

- Nenadzirano predtreniranje (*engl. Unsupervised Pre-training*):

Tijekom ove faze, GPT se izlaže masivnom, nelabeliranom tekstualnom korpusu, učeći obrasce, strukture i jezične veze iz podataka. Model koristi jednosmjerni, autoregresivni pristup, gdje predviđa sljedeću riječ u nizu na temelju prethodnih riječi [1]. Ovaj pristup omogućuje GPT-u učenje sekvencijalnih ovisnosti i konteksta, što je ključno za generiranje smislenog teksta. Jednosmjerno treniranje fokusira se samo na kontekst lijevo od značajne riječi, što znači da se svaka riječ u nizu predviđa isključivo na temelju prethodnih riječi. Ovaj pristup omogućuje GPT-u izvršavanje zadataka koji zahtijevaju generiranje teksta, kao što su kreativno pisanje ili dijalози. Međutim, taj isti jednosmjerni pristup može biti i ograničenje kada su u pitanju zadaci koji zahtijevaju razumijevanje ili sažimanje informacija iz cijelog teksta.

- Nadzirano fino podešavanje (*engl. Supervised Fine-tuning*):

Nakon prethodnog obučavanja, GPT prolazi fazu nadziranog finog podešavanja, gdje se model usavršava koristeći manji, specifičan skup podataka sa označenim primjerima. Tijekom finog podešavanja, model se trenira za obavljanje specifičnih zadataka, kao što

su prevođenje, sažimanje ili odgovaranje na pitanja, prilagođavajući svoje težine na temelju pruženih primjera. Ova faza je ključna za prilagodbu općeg modela, stvorenog tijekom prethodnog obučavanja, kako bi izvrsno obavljao specifične primjene. Učinkovitost procesa finog podešavanja uvelike ovisi o kvaliteti i količini označenih podataka, kao i o relevantnosti zadataka specifičnih primjera za namjeravanu primjenu.

### 2.7.2. BERT metodologija treniranja

BERT koristi drugačiji pristup treniranju koji koristi dvosmjerno učenje, omogućujući modelu hvatanje dubljeg razumijevanja konteksta i semantike unutar teksta.

- Maskirano jezično modeliranje (*engl. Masked Language Modeling (MLM)*) [12]:

U fazi predtreniranja, BERT koristi metodu poznatu kao Maskirano jezično modeliranje. Tijekom ovog procesa, određeni postotak riječi u svakoj rečenici nasumično se maskira (skriva), a model se trenira da predvidi ove maskirane riječi na temelju konteksta koji pružaju okolne riječi. Za razliku od GPT-ovog jednosmjernog pristupa, BERT-ovo treniranje je dvosmjerno, što znači da razmatra cijeli kontekst—i riječi prije i nakon maskirane riječi. Ovaj dvosmjerni pristup omogućuje BERT-u da razvije sveobuhvatnije razumijevanje teksta, razumijevajući odnose i nijanse koje bi mogle biti propuštene u jednosmjernom modelu. Ova metodologija treniranja čini BERT-a izuzetno učinkovitim za zadatke koji zahtijevaju duboko razumijevanje jezika, kao što su analiza sentimenta, prepoznavanje entiteta i strojno razumijevanje.

- Predviđanje sljedeće rečenice (*engl. Next Sentence Prediction (NSP)*) [12]:

Drugi ključni element BERT-ovog treninga je metoda predviđanja sljedeće rečenice. Tijekom ove faze, model se trenira da razumije odnose između rečenica predviđajući je li određena rečenica logički sljedeća nakon druge. To se postiže predstavljanjem modela s parovima rečenica tijekom treninga: neki parovi su uzastopne rečenice iz teksta, dok su drugi nasumično odabrane, razbacane rečenice. Model uči razlikovati ove dvije situacije, poboljšavajući svoju sposobnost za izvođenje zadataka koji uključuju razumijevanje na razini rečenice, kao što su klasifikacija dokumenata, odgovaranje na pitanja i utvrđivanje teksta. Predviđanje sljedeće rečenice posebno je koristan oblik treninga za poboljšanje performansi modela na zadacima gdje je kritično razumijevanje slijeda i koherencije rečenica.

Različite metodologije treniranja GPT-a i BERT-a rezultiraju modelima koji su prikladni za različite vrste zadataka obrade prirodnog jezika. GPT-ov jednosmjerni trening i autoregresivna priroda čine ga izvrsnim izborom za zadatke koji uključuju generiranje novog sadržaja, kao što su kreativno pisanje ili generiranje dijaloga. S druge strane, BERT-ov dvosmjerni trening i fokus na razumijevanju konteksta čine ga pogodnijim za zadatke koji zahtijevaju duboko razumijevanje jezika, kao što su klasifikacija teksta, pretraga informacija i odgovaranje na pitanja na temelju konteksta. Ove razlike naglašavaju važnost odabira odgovarajućeg modela na temelju specifičnih zahtjeva zadatka. Dok GPT-ove generativne sposobnosti dolaze do izražaja u stvaranju tečnog i koherentnog teksta, BERT-ova snaga leži u njegovoj sposobnosti razumijevanja i interpretiranja složenih jezičnih obrazaca, čineći ga neophodnim za zadatke koji zahtijevaju detaljnu analizu i razumijevanje teksta.

### 3. TRANSFORMERI

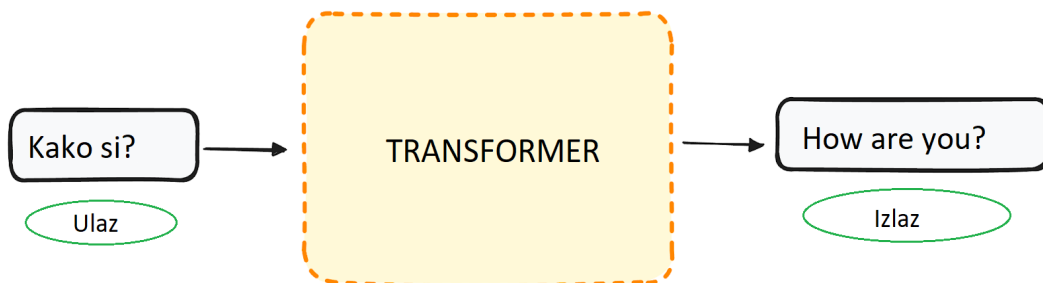
#### 3.1. Uvod u transformere

Transformeri su arhitekturni model koji je postao temeljni alat u području obrade prirodnog jezika. Ova inovativna tehnika temelji se na mehanizmu pažnje (engl. *attention mechanism*) i omogućuje modelima da učinkovito razumiju i generiraju jezične strukture. Transformeri su prvi put predstavljani u znanstvenom radu "Attention is All You Need", Vaswani et al [1].

#### 3.2. Arhitektura transformera

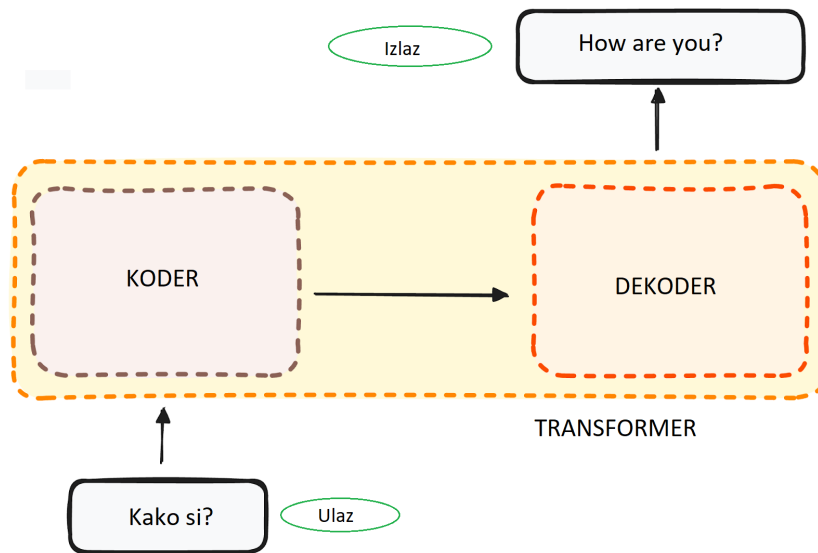
Originalno zamišljeni za neuronski prijevod, transformeri [36] su izvrsni u pretvaranju ulaznih nizova u izlazne nizove. To je prvi model koji se u potpunosti oslanja na mehanizam samo-pažnje za izračunavanje reprezentacija svojih ulaza i izlaza, ne koristeći sekvencijalno poravnate rekurentne neuronske mreže (engl. Recurrent Neural Network, RNN) [37]. Glavna karakteristika arhitekture transformera je što zadržavaju model kodera-dekoder.

Ako transformer za strojno prevođenje promatramo kao jednostavnu crnu kutiju, on bi kao ulaz primao rečenicu na jednom jeziku, primjerice hrvatskom, te bi kao izlaz davao njezin prijevod na drugom jeziku, primjerice engleskom.



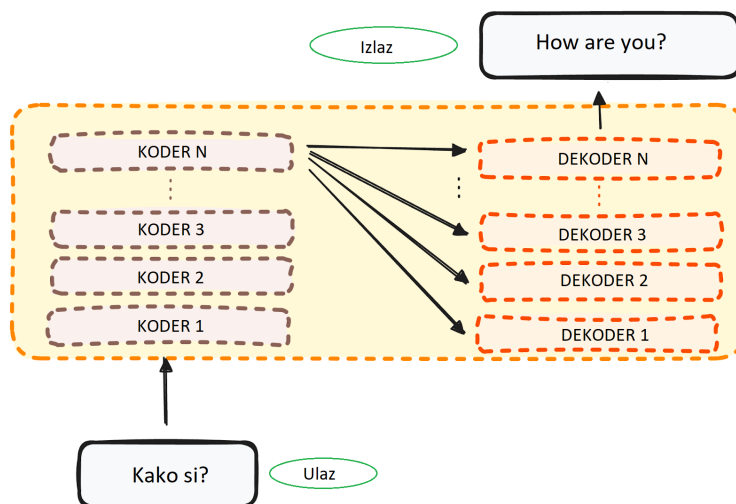
Ako se pogleda slika 3.1, može se vidjeti kako je transformer sastavljen od dva glavna dijela:

- Koder koji uzima ulazni niz te izbacuje matricu reprezentaciju tog ulaza. Na primjer, hrvatska rečenica „Kako si?“
- Dekoder uzima tu kodiranu reprezentaciju i iterativno generira izlazni niz. U ovom primjeru, prevedena rečenica je „How are you?“



Sl.3.1. Struktura kodera-dekodera [36].

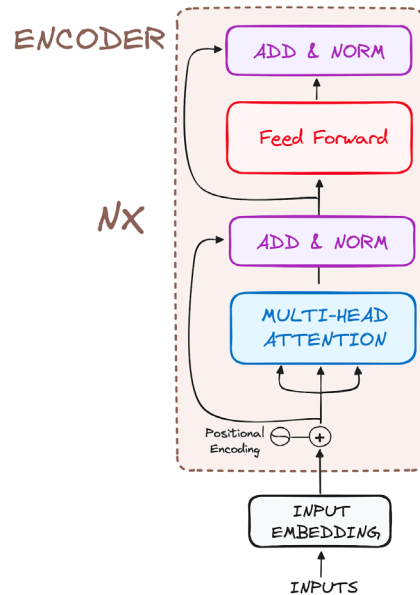
Koder i dekodeer u transformer modelu su zapravo višeslojni stogovi s jednakim brojem slojeva za svaki. Struktura kodera je konzistentna u svim slojevima, pri čemu se izlaz jednog sloja prenosi u sljedeći. Dekoderi također imaju identičnu strukturu, primajući ulaz iz zadnjeg koderskog sloja ili prethodnog dekoderskog sloja. Broj slojeva u transformer modelu može se replicirati prema potrebi, što je ilustrirano na slici 3.2.



Sl.3.2. Višeslojevita struktura kodera-dekodera [36].

## Radni tok koderera

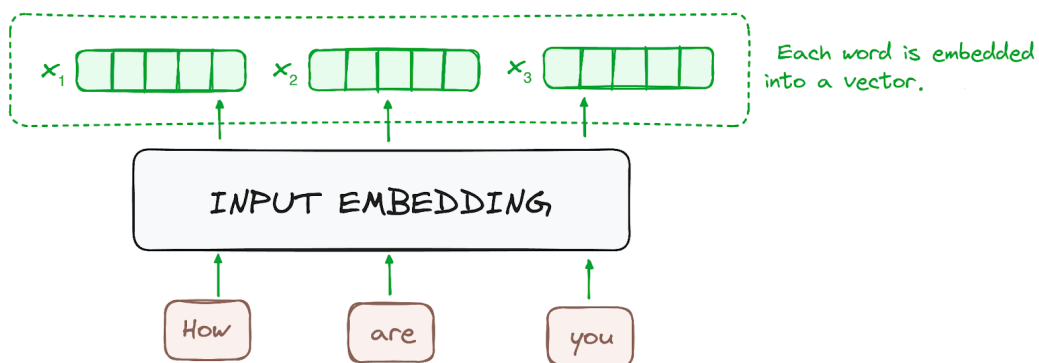
Koder je temeljna komponenta arhitekture transformera. Osnovna funkcija koderera je transformirati ulazne tokene u kontekstualizirane reprezentacije. Koder transformera uzima kontekst svakog tokena u odnosu na cijeli niz.



Sl.3.3. Struktura koderera [36].

## Ulazne ugradnje (engl. *Input embeddings*)

Ulazne ugradnje [36] događaju se samo na najnižem sloju koderera. Koder započinje pretvaranjem ulaznih tokena - riječi ili podriječji - u vektore koristeći slojeve ugradnje. Te ugradnje sadrže semantičko značenje tokena i pretvaraju ih u numeričke vektore. Svi koderi primaju listu vektora, što je vidljivo na slici 3.4. Na najnižem sloju koderera to bi bile ugradnje riječi, dok bi u ostalim slojevima koderima to bio izlaz iz sloja koderera koji se nalazi neposredno ispod njih.

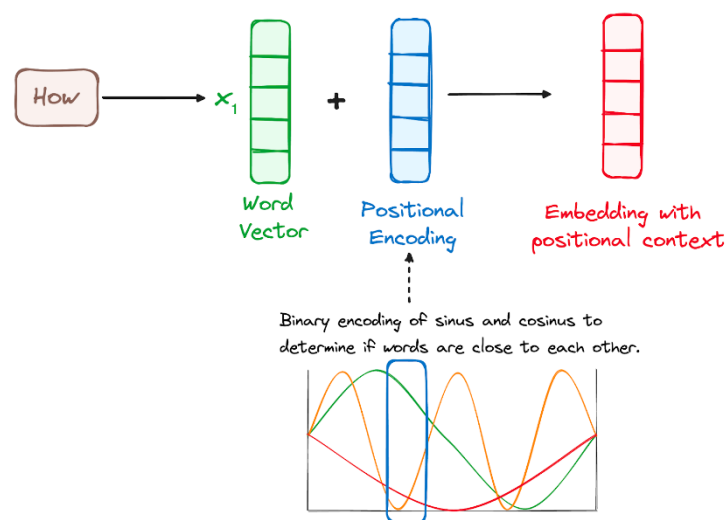


Sl.3.4. Struktura koderera – ulazne ugradnje [36].

## Pozicijsko kodiranje (engl. Positional encoding)

Budući da transformeri nemaju mehanizam povratne veze kao RNN-ovi, koriste pozicijska kodiranja [36] dodana ulaznim ugradnjama kako bi pružili informacije o poziciji svakog tokena u nizu. To im omogućuje razumijevanje položaja svake riječi unutar rečenice. Kako bi to postigli, koriste kombinacije različitih sinus funkcija [38] za stvaranje pozicijskih vektora koji omogućavaju korištenje ovog pozicijskog koda za rečenice bilo koje duljine.

U ovom pristupu svaka dimenzija je predstavljena jedinstvenim frekvencijama i pomacima vala, s vrijednostima koje se kreću od -1 do 1, učinkovito predstavljajući svaku poziciju što je prikazano na slici 3.5.



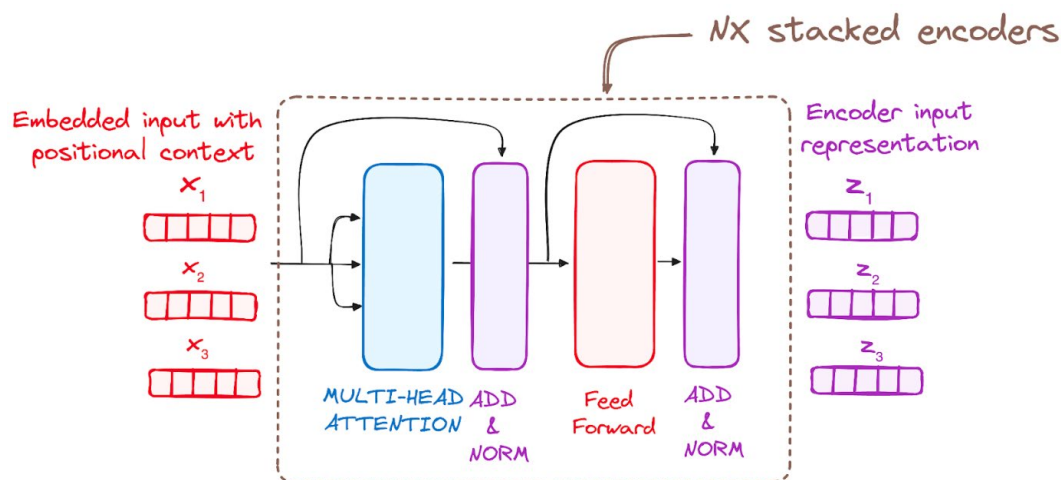
Sl.3.5. Struktura koda – pozicijsko kodiranje [36].

## Stog koderskih slojeva

Transformerski koder se sastoji od stoga identičnih slojeva. Koderski sloj služi za transformaciju svih ulaznih sekvenci u kontinuiranu apstraktnu reprezentaciju koja obuhvaća naučene informacije iz cijele sekvence. Ovaj sloj sadrži dva pod-modula:

- Mehanički mehanizam s više glava (*engl. mutli-head attention*).
- Potpuno povezana unaprijedna mreža (*engl. feed-forward network*).

Dodatno, uključuje rezidualne veze oko svakog pod-sloja nakon kojeg slijedi sloj normalizacije.



Sl.3.6. Struktura koderu – stog koderskih slojeva [36].

### Mehanizam samo-pažnje s više glava (*engl. Multi-Headed Self-Attention Mechanism*)

U koderu, mehanizam s više glava koristi specijalizirani mehanizam pažnje poznat kao samo-pažnja (*engl. self-attention*). Ovaj pristup omogućava modelima da povežu svaku riječ u ulazu s ostalim riječima. Na primjer, u danom primjeru, model može naučiti povezati riječ "kako" s "si". Ovaj mehanizam omogućava koderu da se fokusira na različite dijelove ulazne sekvence dok obrađuje svaki token. Mehanizam izračunava [1] rezultate pažnje temeljene na:

- Upitu - vektor koji predstavlja specifičnu riječ ili token iz ulazne sekvence u mehanizmu pažnje.
- Ključu - vektor u mehanizmu pažnje koji odgovara svakoj riječi ili tokenu u ulaznoj sekvenci.
- Vektorska vrijednost - povezana je s ključem i koristi se za konstrukciju izlaza sloja pažnje. Kada se upit i ključ dobro podudaraju, što u prijevodu znači da imaju visok rezultat mehanizma pažnje, odgovarajuća vrijednost je naglašena na izlazu.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1-1)$$

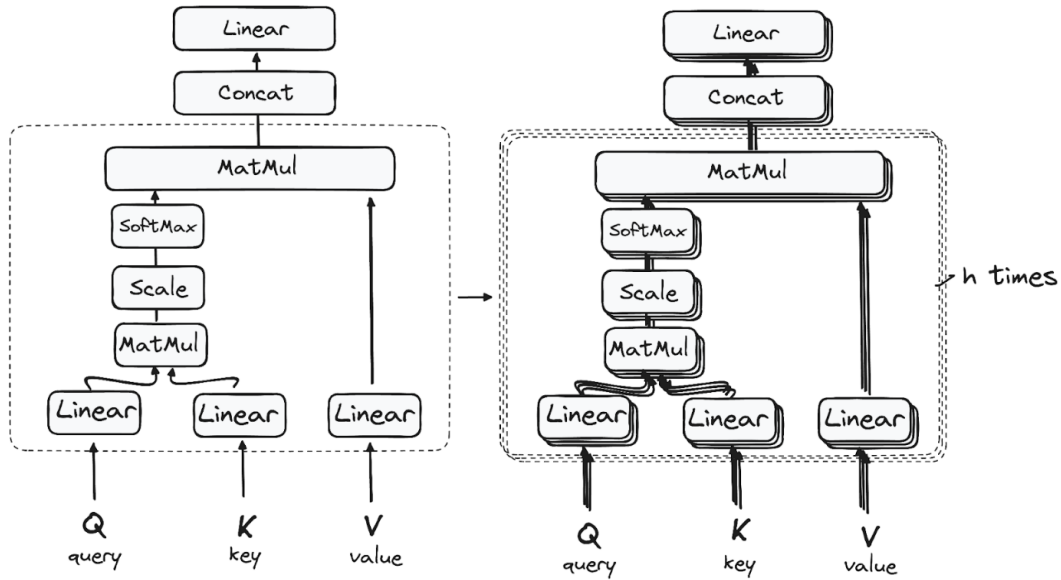
Gdje su  $Q$ ,  $K$  i  $V$  matrice upita (Upit), ključeva (Ključ) i vrijednosti (Vektorska vrijednost) koje se dobivaju linearnom transformacijom ulaznih vektora.

Ovaj prvi modul samo-pažnje omogućava modelu da razumije kontekstualne informacije iz cijele sekvence. Umjesto izvođenja jedne funkcije pažnje, upiti, ključevi te vektorske vrijednosti



su linearno projicirani  $h$  puta. Na svakoj od ovih projiciranih verzija upita, ključeva i vektorskih vrijednosti, mehanizam pažnje se izvodi paralelno, dajući  $h$ -dimenzionalne izlazne vrijednosti.

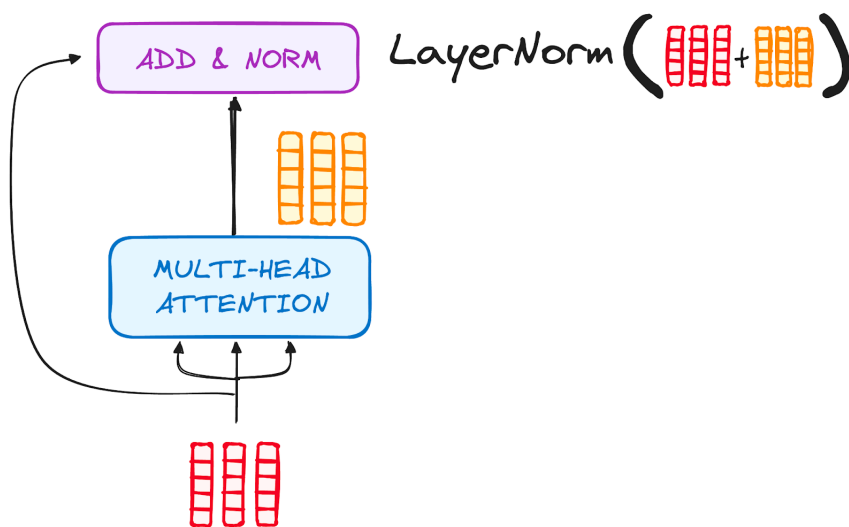
Detaljna arhitektura prikazana je na donjoj slici 3.7.



Sl.3.7. Struktura mehanizma samopažnje s više glava [36].

### Normalizacija i rezidualne veze

Svaki pod-sloj u sloju kodera prati korak normalizacije [36]. Također, izlaz svakog pod-sloja dodaje se njegovom ulazu (rezidualna veza) kako bi se pomoglo ublažiti problem nestajanja gradijenta, što omogućava veće modele. Ovaj proces će se ponoviti i nakon unaprijedne mreže.



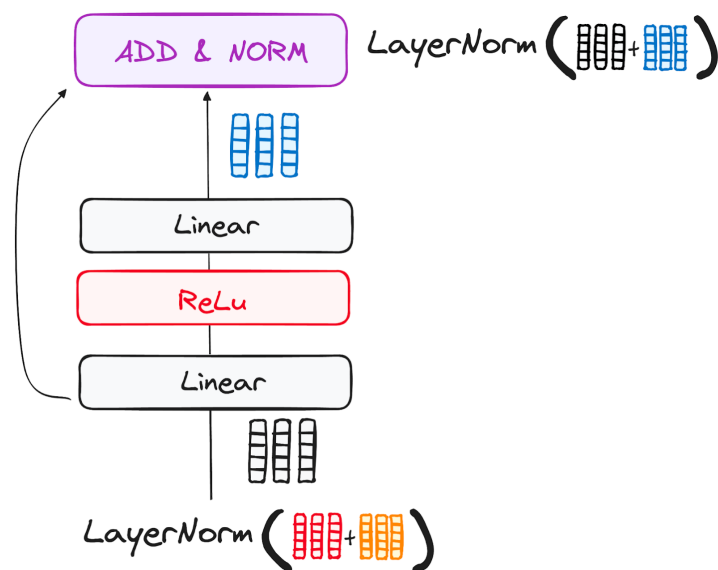
Sl.3.8. Struktura kodera - normalizacija i rezidualna veza nakon mehanizma pažnje s više glava [36].

## Unaprijedna mreža

Put normaliziranog rezidualnog izlaza nastavlja se dok ne prođe kroz duboku unaprijednu mrežu, ključnu fazu za dodatno pročišćavanje. Nakon što su vrijednosti transformirane pomoću mehanizma pažnje, one prolaze kroz unaprijednu neuronsku mrežu koja se sastoji od dva linearna sloja s funkcijom aktivacije [1] između njih.

- **Prvi linearni sloj (Linear 1):** Ulazni vektor se množi matricom težina i dodaje se prag. Rezultat prolazi kroz funkciju aktivacije (ReLU [39]).
- **Drugi linearni sloj (Linear 2):** Izlaz iz prvog sloja se ponovo množi drugom matricom težine i dodaje se prag.

Unaprijedne neuronske mreže u transformerima služe za daljnju obradu i transformaciju informacija koje su već prošle kroz mehanizam pažnje. One dodaju nelinearnost i omogućavaju mreži da uči složene uzorke i odnose unutar podataka.



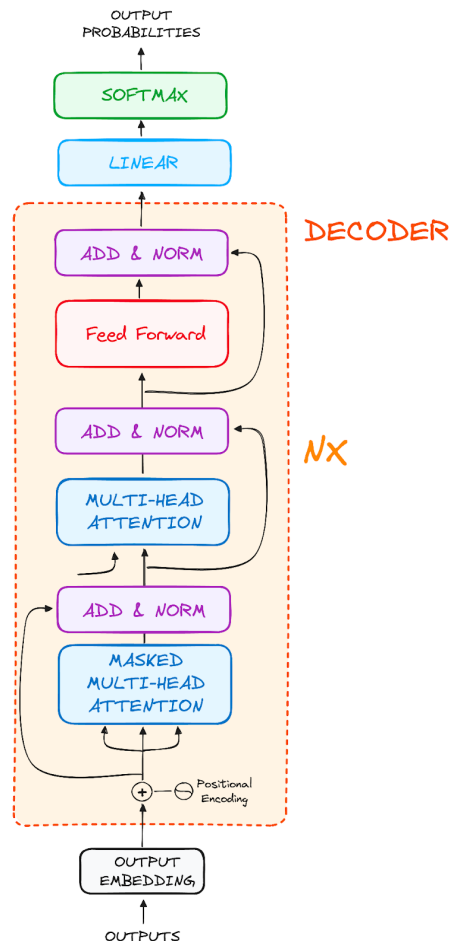
Sl.3.9. Struktura kodera - pod-sloj neuronske unaprijedne mreže [36].

## Izlaz kodera

Izlaz kodera je niz vektora  $[h_1, h_2, \dots, h_n]$  gdje svaki vektor  $h_i$  predstavlja kontekstualnu reprezentaciju tokena  $x_i$  iz ulazne sekvence. Ovi vektori sadrže informacije o svakom tokenu, uzimajući u obzir njegov odnos prema ostalim tokenima u sekvenci.

## Radni tok dekodera

Uloga dekodera usredotočena je na stvaranje tekstualnih sekvenci. Odražavajući koder, dekodер je opremljen sličnim setom pod-slojeva. Ima dva sloja pažnje s više glava, unaprijednu mrežu i uključuje rezidualne veze te normalizaciju sloja nakon svakog pod-sloja.



Sl.3.10. Struktura dekodera - Globalna struktura dekodera [36].

Ove komponente funkcioniraju na način sličan slojevima koder, ali sa zapletom: svaki sloj pažnje s više glava u dekodерu ima svoj jedinstveni zadatak. Krajnji proces dekodera uključuje linearni sloj koji služi kao klasifikator, završavajući sa softmax funkcijom za izračunavanje vjerojatnosti različitih riječi. Struktura transformerskog dekodera posebno je dizajnirana da generira ovaj izlaz dekodiranjem kodiranih informacija korak po korak. Važno je primijetiti da dekodер radi na autoregresivan način, započinjući svoj proces s početnim tokenom. On koristi listu prethodno generiranih izlaza kao svoje ulaze zajedno s izlazima iz koder koji sadrže informacije mehanizma pažnje iz početnog ulaza. Ovaj proces dekodiranja nastavlja dok dekodер ne dođe do ključnog trenutka: generacija tokena koji signalizira kraj izlaza.

## Izlazne ugradnje

Na početnoj liniji dekodera, proces se kopira od kodera. U ovom procesu, ulaz prvo prolazi kroz sloj ugradnje.

## Pozicijsko kodiranje

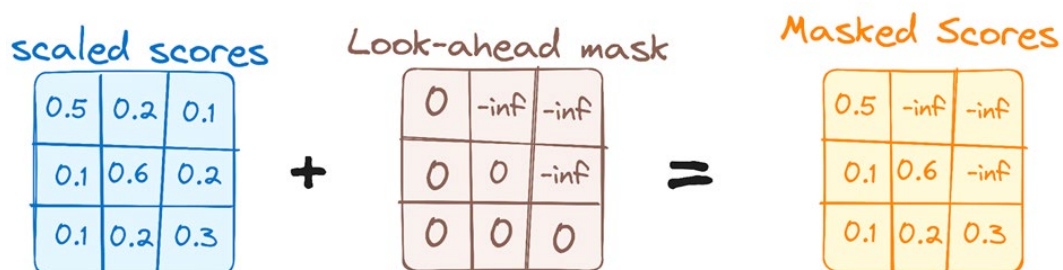
Nakon izlazne ugradnje, ulaz prolazi sloj pozicijskog kodiranja. Ovaj niz je dizajniran kako bi odradio pozicijske ugradnje. Te pozicijske ugradnje zatim se šalju u prvi sloj mehanizma pažnje dekodera s više glava gdje se izračunavaju rezultati pažnje specifični za ulaz dekodera.

## Stog dekoderskih slojeva

Dekoder se sastoji od stoga identičnih slojeva (6 u originalnom transformerskom modelu). Svaki sloj ima tri glavne [1] pod-komponente:

## Maskirani mehanički mehanizam samo-pažnje

Ovo je slično mehanizmu samo-pažnje u koderu, ali s ključnom razlikom: sprječava pozicije da prisustvuju budućim pozicijama, što znači da svaka riječ u sekvenci nije pod utjecajem budućih tokena. Na primjer, kada se izračunavaju rezultati pažnje za riječ "si", važno je da "si" nije dio riječi "kako", što je buduća riječ u sekvenci.



Sl.3.11. Radni tok dekodera - Prva maska višestruke pažnje [36].

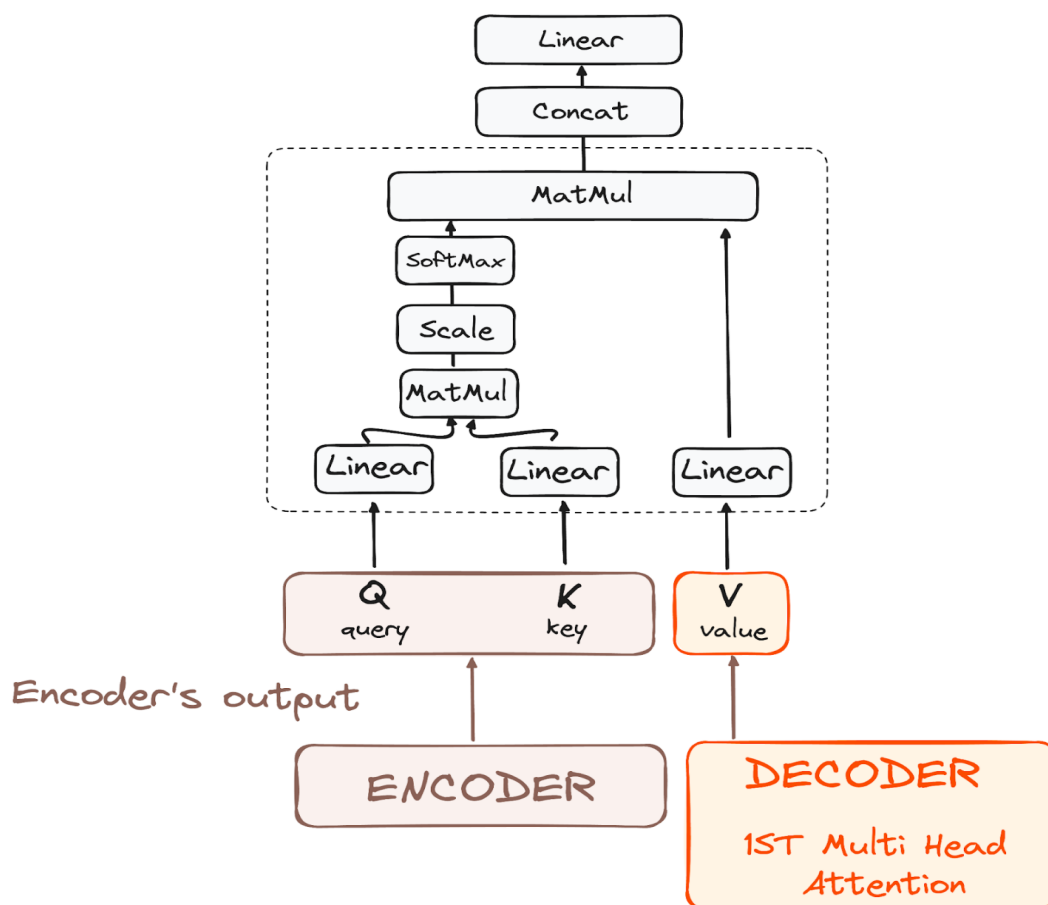
Ovo maskiranje osigurava da predviđanja za određenu poziciju mogu ovisiti samo o poznatim izlazima na pozicijama prije nje.

## Višeglavni mehanički mehanizam pažnje kodera-dekoder ili križna pažnja

U drugom višeglavom sloju pažnje dekoder vidimo jedinstvenu interakciju između komponenti kodera i dekoder. Ovdje izlazi iz kodera preuzimajući uloge i upita i ključeva dok izlazi iz prvog višeglavog sloja pažnje dekoder služe kao vrijednosti.

Ovaj setup učinkovito usklađuje ulaz kodera s ulazom dekoder, osnažujući dekoder da identificira i naglasi najrelevantnije dijelove ulaza kodera.

Nakon toga, izlaz iz ovog drugog sloja višeglavog mehanizma pažnje zatim je pročišćen kroz unaprijednu mrežu, dodatno poboljšavajući obradu.



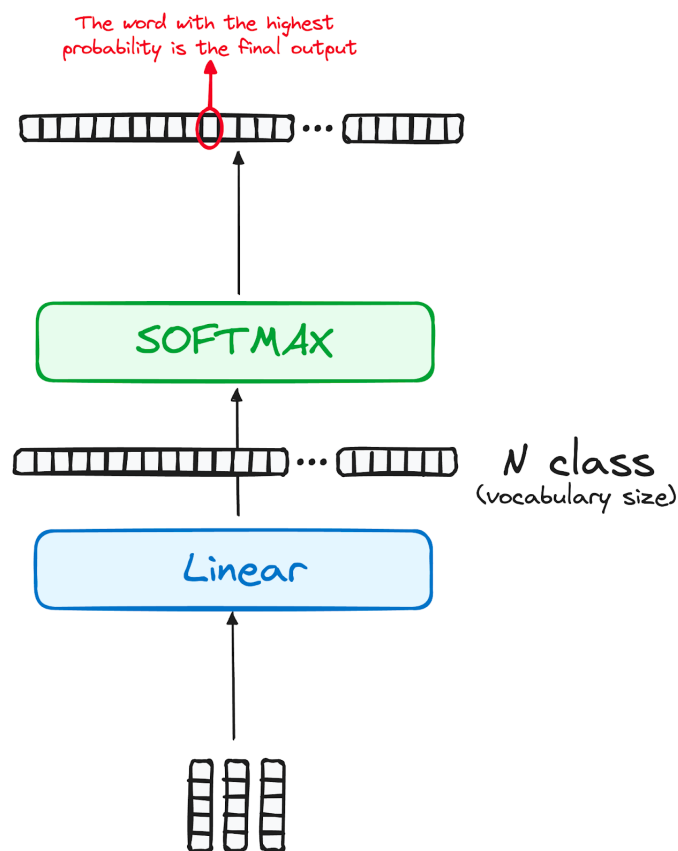
Sl.3.12. Radni tok dekoder - Pažnja kodera-dekoder [36].

## Unaprijedna mreža

Slično kao kod kodera, svaki sloj dekoder uključuje potpuno povezanu unaprijednu mrežu primijenjenu na svaku poziciju zasebno i identično.

## Linearni klasifikator i softmax za generiranje vjerojatnosti izlaza

Put podataka kroz transformerski model kulminira njegovim prolaskom kroz konačni linearni sloj koji funkcionira kao klasifikator. Veličina ovog klasifikatora odgovara ukupnom broju klasa uključenih (broj riječi sadržanih u vokabularu). Na primjer, u scenariju s 1000 različitih klasa koje predstavljaju 1000 različitih riječi, izlaz klasifikatora će biti niz s 1000 elemenata. Ovaj izlaz zatim se uvodi u sloj softmax koji ga transformira u raspon vjerojatnosnih rezultata, svaki ležeći između 0 i 1. Najviši od ovih rezultata vjerojatnosti je ključan; njegov odgovarajući indeks izravno upućuje na riječ koju model predviđa kao sljedeću u sekvenci.



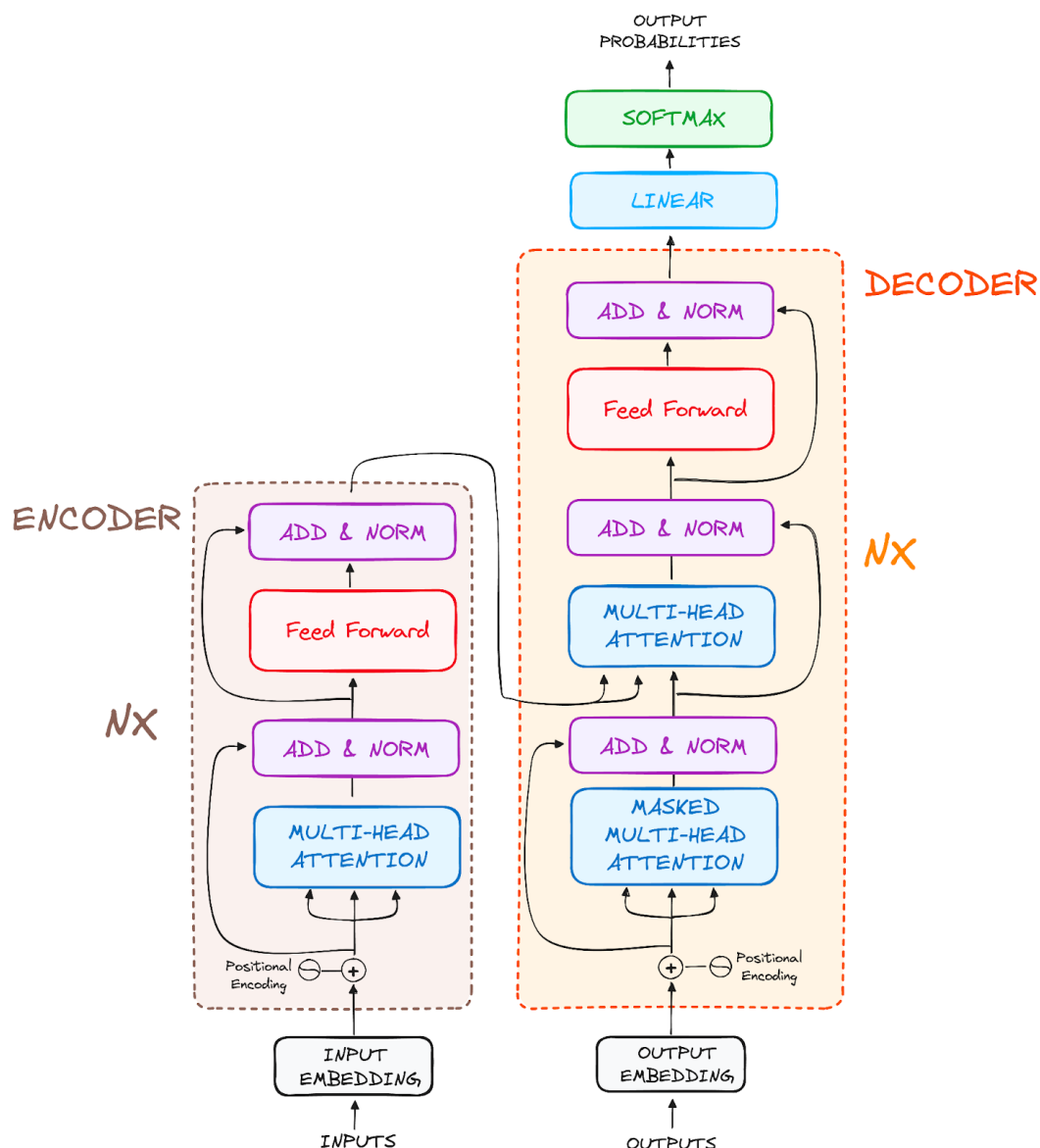
Sl.3.13. Radni tok dekodera - Konačni izlaz transformera [36].

## Normalizacija i rezidualne veze

Svaki pod-sloj (maskirani samo-pažnja, križna pažnja, unaprijedna mreža) prati korak normalizacije i uključuje rezidualnu vezu oko njega.

## **Izlaz dekodera**

Izlaz konačnog sloja dekodera transformira se u predviđenu sekvencu, tipično kroz linearni sloj praćen softmaxom za generiranje vjerojatnosti preko vokabulara. Dekoder u svom operativnom toku uključuje svježe generirani izlaz u svoj rastući popis ulaza, a zatim nastavlja s procesom dekodiranja. Ovaj ciklus se ponavlja dok model ne predvidi specifični token koji signalizira završetak. Token predviđen s najvećom vjerojatnošću dodjeljuje se kao zaključna klasa, često predstavljena tokenom završetka. Dekoder nije ograničen na jedan sloj. Može biti strukturiran s N slojeva, svaki se gradi na ulazu primljenom od koda i njegovih prethodnih slojeva. Ova slojevita arhitektura omogućava modelu da proširi svoj fokus i izvuče različite uzorke pažnje preko svojih glava pažnje. Takav višeslojni pristup može znatno poboljšati sposobnost modela za predviđanje jer razvija složenije razumijevanje različitih kombinacija pažnje. Konačna arhitektura je prikazana na slici 3.14 (iz originalnog rada “Attention is All You Need”).



Sl.3.14. Struktura transformera [36].

### 3.3. Primjena transformera u obradi prirodnog jezika

Transformeri su se pokazali kao iznimno korisnima u brojnim područjima obrade jezika. Jedan od najboljih primjera je strojno prevođenje, gdje transformeri postižu odlične rezultate. Modeli poput "Google Neural Machine Translation" (GNMT) koriste transformere za prevoditeljske zadatke i omogućuju kvalitetno prevođenje između različitih jezika [40]. Također, transformeri se koriste za generiranje sažetaka tekstova, prepoznavanje govora, analizu sadržaja na društvenim mrežama, chatbotove te pretraživanje informacija.

Izraz "unaprijed obučeni" ili "predtreniranje" odnosi se na modele koji su prethodno obučeni na velikim skupovima podataka, koji obično sadrže milijune dokumenata koji pokrivaju širok



raspon tema. Ovo izlaganje omogućuje modelima da nauče suptilne obrasce jezika, uključujući sintaksu, semantiku i kontekst. Korištenjem ovih unaprijed obučениh modela, virtualno stječete mogućnost iskorištavanja njihovog naučenog znanja, eliminirajući potrebu za obukom modela od nule. Ovaj pristup štedi značajne računalne resurse i vrijeme, jer je sada samo potrebno uvježbati model za svoj specifični zadatak. Na primjer, korištenje označenog skupa podataka koji se sastoji od parova rečenica i oznaka neutralnih, negativnih ili pozitivnih za izvođenje analize raspoložnja.

Transformeri su postali neizostavan dio modernog strojnog učenja. Konkretno, oni nude neusporedivu sposobnost u obradi prirodnog jezika, ali njihov potencijal seže čak i izvan ovog područja. Kako su transformeri pokazali obećavajuće rezultate u domenama kao što su računalni vid i prepoznavanje zvuka, nagovještavajući budućnost u kojoj bi mogli postati arhitektura strojnog učenja opće namjene. Međutim, prava prednost transformera leži u njihovoj dostupnosti.

Osim toga, porast tehnika inženjeringa upita s nula pokušaja (*engl. zero-shot*) i s nekoliko pokušaja (*engl. few-shot*) proširio je primjenjivost transformera. Tehnika inženjeringa upita s nula (*eng. zero-shot*) odnosi se na sposobnost modela da se nosi sa zadacima za koje nije posebno obučен. U biti, on može razumjeti i obavljati zadatke bez presedana na temelju svoje opsežne obuke. S druge strane, kratkotrajno učenje dokazuje kako model može brzo naučiti izvršavati nove zadatke nakon obuke na vrlo maloj količini podataka povezanih s tim zadatkom. Ova tehnika koristi postojeće znanje modela iz njegove opsežne obuke. Ove napredne tehnike, u kombinaciji s resursima, čine transformere prvim izborom za širok raspon zadataka obrade prirodnog jezika.

Međutim, koliko god ovi različiti modeli transformera bili impresivni, dolaze s određenim ograničenjima. Ako uzmemo u obzir takozvane modele s milijardama parametara, to se odnosi na skalu i parametre koji se mogu naučiti ili "težine" ovih modela transformera. Na primjer, rani modeli transformera imali su 110 milijuna parametara, ali kako se napredak u ovom području nastavlja, to je dovelo do još većih modela sa 175 milijardi parametara i više. Ovi "modeli s milijardama parametara" pokazuju impresivnu sposobnost generiranja koherentnih i kontekstualno relevantnih odgovora u razgovoru. Međutim, unatoč veličini i impresivnoj izvedbi, ovi modeli imaju svoja ograničenja. Njihova praktična primjenjivost može opasti u određenim specijaliziranim područjima kao što su financije ili zdravstvo, gdje je specifičan kontekst bitan. Računalni i memorijski zahtjevi ovih modela mogu njihovu implementaciju

iznenada učiniti posebno izazovnom, posebno u stvarnom vremenu, gdje su brzi i točni rezultati ključni. Međutim, potencijal ovih različitih varijacija modela transformera ne može se podcijeniti. Oni pomiču granice onoga što je moguće u razumijevanju i generiranju prirodnog jezika, klasifikaciji teksta, prevođenju i mnogim drugim područjima. Ali kao i kod svakog drugog alata, ključ leži u uravnoteženom pristupu specifičnom problemu.

### **3.4. Prednosti i ograničenja transformera**

Transformeri donose niz prednosti u obradi prirodnog jezika. Njihova sposobnost modeliranja konteksta i razumijevanja jezičnih struktura omogućuje im generiranje kvalitetnih i relevantnih tekstova. Osim toga, transformeri su dobri u učenju u par pokušaja (engl. *few-shot learners*), što znači da mogu naučiti zadatke s malo primjera, što je Brown et al. (2020) [16] dokazao u istraživanju. Ova sposobnost čini transformere fleksibilnim alatima koji se mogu koristiti u raznim zadacima obrade prirodnog jezika.

Međutim, transformeri su skloni prenaučavanju (engl. *overfitting*) ako nemaju dovoljno podataka ili nisu pravilno regulirani. Ova ograničenja mogu utjecati na njihovu sposobnost generalizacije na novim ili specifičnim jezičnim zadacima.

Jedan od nedostataka je potencijalno preklapanje testnih podataka sa validacijskim podacima, što može povećati percepciju izvedbe modela. Osim toga, modeli opće namjene poput GPT-4 i GPT-4o mogu se nositi sa zadacima koji zahtijevaju dvosmjerni kontekst, ograničavajući njihovu učinkovitost u određenim slučajevima upotrebe.

Specijalizirane domene poput financija ili zdravstva mogu zahtijevati dodatno podešavanje kontekstualnog razumijevanja, što možda nije svojstveno modelima opće namjene kao što je GPT-3. Štoviše, računalni i memorijski zahtjevi ovih modela mogu njihovu implementaciju učiniti izazovnom, posebno u sustavima u stvarnom vremenu. Klasični modeli posebno dizajnirani i optimizirani za specifične zadatke mogu nadmašiti ove velike modele u smislu brzine, potrošnje resursa i točnosti u određenim kontekstima. Stoga je neophodan uravnotežen pristup specifičan za pojedini problem prilikom odabira modela, shvaćajući da veće nije uvijek bolje.

Postojeći primjeri primjene transformera u obradi prirodnog jezika svjedoče o njihovoj izvanrednoj učinkovitosti. Primjerice, GPT-4o (Generative Pre-trained Transformer 4o) model, razvijen od strane OpenAI, je jedan od najmoćnijih modela obrade prirodnog jezika do sada. GPT-4o je prethodno treniran na velikom skupu podataka iz interneta i može izvesti različite

zadatke kao što su prevoditeljski zadaci, generiranje teksta, odgovaranje na pitanja i još mnogo toga [41]. Ovaj model je postigao impresivne rezultate u generiranju koherentnih i prirodnih tekstova.

Transformeri su promijenili način obrade prirodnog jezika i otvorili vrata brojnim inovativnim aplikacijama. Njihova sposobnost razumijevanja i generiranja jezičnih struktura donosi nove mogućnosti za komunikaciju ljudi s računalima na ljudima prirodan način. Unatoč nekim ograničenjima, transformeri se neprekidno razvijaju i poboljšavaju s napretkom tehnologije i sve većom dostupnošću jezičnih podataka. Kao rezultat toga, realno je za očekivati kako će transformeri nastaviti oblikovati budućnost obrade prirodnog jezika i doprinijeti razvoju novih tehnologija koje će transformirati način na koji koristimo jezik u digitalnom okruženju.

## 4. MEHANIZAM PAŽNJE U STROJNOM UČENJU (*ENGL. ATTENTION MECHANISM IN MACHINE LEARNING*)

Mehanizam pažnje u modelima transformera ključan je element koji omogućuje modelu da se fokusira na različite dijelove ulazne sekvence prilikom generiranja izlazne sekvence.

Mehanizam pažnje omogućuje svakom elementu u ulaznoj sekvenci da usmjerava pažnju na sve pozicije u sekvenci, što modelu omogućuje da odredi koje su pozicije ključne za određeni kontekst. Ovaj proces je konzistentan kroz različite pozicije zbog upotrebe istog skupa parametara. Konkretno, u arhitekturi transformera svaki element (ili riječ) u ulaznoj sekvenci može se usredotočiti na sve pozicije u sekvenci, što modelu omogućuje da odredi koje pozicije su ključne za dati kontekst. To se postiže korištenjem istog skupa parametara, čineći proces konzistentnim kroz različite pozicije. Dakle, mehanizam samopažnje pruža modelu sposobnost da se fokusira na segmente ulazne sekvence koji su najrelevantniji za trenutni korak obrade, osiguravajući da se relevantne kontekstualne informacije zadrže i naglase.

### 4.1. Vrste mehanizama pažnje

Postoji nekoliko vrsta mehanizama pažnje koji se koriste u NLP modelima:

1. **Mehanizam skaliranja unutarnjeg produkta (*engl. Scaled Dot-Product Attention*)** [42]:

Ovo je osnovni mehanizam pažnje korišten u arhitekturi transformera. Računa se tako da se uzima skalarni umnožak vektora upita (*engl. query*) i ključa (*engl. key*), a zatim se skalira kvadratnim korijenom dimenzionalnosti kako bi se izbjegle prevelike vrijednosti. Dobiveni rezultati prolaze kroz softmax funkciju kako bi se izračunale težine pažnje, koje se potom množe s vektorima vrijednosti (*engl. value*) za stvaranje izlaza.

2. **Mehanizam pažnje s više glava (*eng. Multi head attention*)** [1], [43]:

Prošireni mehanizam skaliranja unutarnjeg produkta je gdje više glava pažnje (*engl. attention heads*) uče različite reprezentacije ulaza. Izlaz svake glave se konkatenira i linearno transformira kako bi se dobio konačni izlaz. Time se omogućuje modelu istovremeno usmjeravanje pozornost na različite dijelove ulaza.

3. **Mehanizam samopažnje (engl. *Self-attention*)** [1], [43]:

Poseban slučaj mehanizma pažnje gdje upiti, ključevi i vrijednosti dolaze iz istog izvora. Široko se koristi u transformerima, gdje svaki token u sekvenci obraća pažnju na sve ostale tokene, omogućujući modelu razumijevanje ovisnosti između tokena bez obzira na njihovu udaljenost.

4. **Križana samopažnja (engl. *Cross-Attention*)** [43], [44]:

Koristi se u modelima sekvenca-sekvencija kao što su modeli za strojno prevođenje. Upiti dolaze iz dekodera, a ključevi i vrijednosti iz enkodera, čime dekodirer može usmjeriti pozornost na različite dijelove ulazne sekvence prilikom generiranja izlazne sekvence.

## 4.2. Vrste prozora pažnje (engl. *Attention window*)

Prozori pažnje su strategije za upravljanje računalnim troškom i upotrebom memorije kod mehanizama pažnje, posebno kod dugih sekvenci.

1. **Klizni prozor pozornosti (engl. *Sliding window attention*)** [45]:

Umjesto da obraća pažnju na sve tokene u sekvenci, model obraća pažnju samo na fiksni prozor okolnih tokena. Prozor se pomiče kroz ulaznu sekvenciju, omogućujući modelu fokus na lokalne kontekste. Ovo smanjuje računalnu složenost s  $O(n^2)$  na  $O(n \cdot w)$ , gdje je  $w$  veličina prozora.

2. **Grupirani upit pozornost (engl. *Grouped query attention (GQA)*)** [46]:

Umjesto da računa pažnju za svaki upit zasebno, GQA grupira više upita zajedno. To omogućuje učinkovito računanje jer više upita može dijeliti dijelove izračuna. GQA je osobito koristan u scenarijima gdje više upita ima preklapajuće distribucije pažnje.

3. **Mehanizam djelomične pažnje (engl. *Sparse attention*):**

Nisu sve pozicije u sekvenci jednako važne. Mehanizam djelomične pažnje obraća pažnju samo na podskup pozicija odabranih prema nekoj heuristici ili naučenom uzorku. Primjeri uključuju Longformer [45], koji kombinira pažnju kliznog prozora s globalnim tokenima koji mogu obratiti pažnju na sve pozicije.

#### 4. Lokalna i globalna pažnja (*engl. Local and global attention*) [47]:

Kombinira lokalne (fiksni prozor) i globalne (cijela sekvenca) mehanizme pažnje kako bi se modelima omogućilo obraćanje pozornosti na važne pozicije globalno, uz održavanje učinkovitosti putem lokalne pažnje. Ovaj pristup koristi se u modelima poput BigBird-a [48] i Longformer-a [45] za balansiranje između kapaciteta modela i učinkovitosti.

### 4.3. Razlike u treniranju bidirekcionalnih i dekodeo-samo modela (BERT vs. GPT)

Razlike između treniranja bidirekcionalnih modela poput BERT-a i dekodeo-samo modela poput GPT-a proizlaze iz njihove arhitekture i predviđenih primjena:

- **BERT (*engl. Bidirectional Encoder Representations from Transformers*)** [12]:
  - Arhitektura: Koristi bidirekcionalni enkoder transformera koji obraća pažnju na sve tokene u sekvenci istovremeno, i s lijeva na desno i s desna na lijevo.
  - Cilj treniranja: Maskirano modeliranje jezika (MLM), gdje su neki tokeni u ulazu maskirani, a model se trenira da ih predvidi. Ovo omogućuje modelu da uhvati kontekst iz oba smjera.
  - Prednosti: Učinkovito hvata duboki bidirekcionalni kontekst, što ga čini vrlo učinkovitim za zadatke poput odgovaranja na pitanja, analize sentimenta i prepoznavanja imenovanih entiteta.
  - Ograničenja: Nije idealan za generativne zadatke jer nije dizajniran za autoregresivno generiranje teksta.
- **GPT (*engl. Generative Pre-trained Transformer*)** [49]:
  - Arhitektura: Koristi jedno-direkcionalni dekodeo transformera koji obraća pažnju na tokene s lijeva na desno, jedan po jedan.
  - Cilj treniranja: Kauzalno modeliranje jezika (*engl. CLM*) [50], gdje se model trenira za predviđanje sljedeće riječi u sekvenci, omogućujući mu generiranje koherentnog i kontekstualno relevantnog teksta.
  - Prednosti: Prikladan za zadatke generiranja teksta poput strojnog prevođenja, dovršavanja teksta i generiranja dijaloga.
  - Ograničenja: Nedostaje mu duboki bidirekcionalni kontekst jer ne može vidjeti buduće tokene.

- **Ključne razlike:**

- Učenje konteksta: BERT uči iz prošlih i budućih konteksta zbog svoje bidirekionalne prirode, dok GPT uči samo iz prošlog konteksta.
- Vrsta zadatka: BERT-ovi zadaci (MLM i NSP) fokusirani su na razumijevanje i klasifikaciju teksta, dok je GPT-ov zadatak (CLM) usmjeren na generiranje teksta.
- Primjene: BERT je bolji za zadatke razumijevanja jezika kao što su odgovaranje na pitanja i analiza sentimenta, dok je GPT bolji za zadatke generiranja teksta i druge kreativne jezične zadatke.

#### 4.4. Jezični modeli i njihove razlike

Jezični modeli temeljeni na dubokom učenju postali su ključni alati u raznim zadacima obrade prirodnog jezika (NLP), kao što su prevođenje, generiranje teksta i dijalog s korisnicima. Različiti modeli imaju specifične svrhe i strategije treniranja, te se razlikuju po načinu na koji se koriste.

- Osnovni modeli (*engl. Causal or Base Models*) [51]:

Osnovni modeli, ili kauzalni modeli su jezični modeli koji generiraju tekst predviđanjem sljedećeg tokena u nizu. Ovi modeli su dizajnirani tako da koriste samo prethodne tokene u nizu kako bi predvidjeli sljedeći token, čime se postiže autoregresivno generiranje teksta.

- Metoda treniranja: Ovi modeli se treniraju koristeći uzročni zadatak modeliranja jezika, gdje je cilj modela predviđanje sljedećeg tokena na temelju niza prethodnih tokena. Na primjer, model bi mogao predvidjeti riječ "mačka" u rečenici "Pas i mačka su".
- Primjena: Ova metoda omogućuje modelima poput GPT-a da budu vrlo učinkoviti u zadacima generiranja teksta, kao što su dopunjavanje teksta, stvaranje dijaloga i strojno prevođenje.
- Inženjering upita: Inženjering upita osnovnih modela uključuje pružanje početnog upita (prompta) iz kojeg model generira nastavak, temeljen na uzročnom modeliranju.

- Instrukcijski modeli (*engl. Instruct models*) [52], [53]:

Instrukcijski modeli su unaprijed trenirani jezični modeli koji su fino podešeni (*engl. fine-tuned*) na temelju posebnih podataka i zadataka kako bi razumjeli i slijedili upute dane u upitu. Ovi modeli su trenirani da slijede eksplicitne upute i pružaju precizne odgovore.

- Metoda treniranja: Instrukcijski modeli su obično trenirani pomoću tehnike poznate kao pojačano učenje iz ljudske povratne informacije (*engl. Reinforcement Learning from Human Feedback (RLHF)*) [54]. U ovoj metodi, ljudski ocjenjivači procjenjuju odgovore modela na niz uputa i rangiraju ih, što model koristi za učenje kako bi pružio točne i korisne odgovore.
- Primjena: Ovi modeli su korisni za zadatke koji zahtijevaju precizne i strukturirane odgovore, kao što su izrada sažetaka, odgovaranje na pitanja i pružanje objašnjenja.
- Inženjering upita: Inženjering upita za instrukcijske modele obično uključuje postavljanje konkretnih pitanja ili zadataka, kao što su "Napiši sažetak za sljedeći članak..." ili "Objasni razliku između...".
- Razgovorni modeli (*engl. Chat models*) [53], [55]:

Razgovorni modeli su specijalizirana verzija instrukcijskog modela koji su fino podešeni za zadatke dijaloga s korisnicima. Oni su dizajnirani da upravljaju višekratnim okvirom dijaloga i održavaju kontekst između uzastopnih razmjena informacija.

- Razlika u odnosu na instrukcijske modele: Iako su razgovorni modeli trenirani slično kao i instrukcijski modeli, oni su dodatno trenirani na bazama podataka koje uključuju konverzijske tokove kako bi mogli održati kontekst dijaloga. Za razliku od instrukcijskog modela, razgovorni modeli ne samo da odgovaraju na pojedinačne upute, već također trebaju održati koherenciju i kontekst u višekratnim dijalozima.
- Primjena: Različiti modeli se koriste u chatbotovima, virtualnim asistentima i drugim aplikacijama koje zahtijevaju prirodan i tečan dijalog s korisnicima.
- Inženjering upita: Inženjering upita za chat modele često uključuje višekratne razmjene s korisnicima, uz održavanje konteksta prethodnih poruka kako bi se omogućio prirodan tok razgovora.

#### **4.5. Skupovi podataka za treniranje**

Trening jezičnih modela ovisi o dostupnosti velikih i raznovrsnih skupova podataka. Neki od najpoznatijih skupova podataka uključuju:

- Alpaca [56]:

Alpaca je fino podešen jezični model. Razvijen je od strane studenata Standroda. Za kreiranje skupa podataka, autori su koristili metodu nadziranog učenja. Počeli su s



postojećim skupom podataka koji sadrži 175 ljudskih uputa i odgovarajućih izlaza. Zatim su koristili model text-davinci-003 kako bi generirali dodatnih 52000 uputa, koristeći se postojećim podacima kao primjerima. Cilj ovog skupa podataka je fino podešavanje modela kako bi precizno slijedili upute korisnika.

- OpenAssistant (LAION) [57]:

OpenAssistant je inicijativa koju vodi LAION (Large-scale Artificial Intelligence Open Network) za stvaranje otvorenog skupa podataka za treniranje modela otvorenog koda. Skup uključuje korpuse razgovora u stilu asistenta, generiran i označen od strane ljudi, koji se sastoji od 161443 poruke raspoređene u 66497 razgovora, na 35 različitih jezika, označenih s 461.292 ocjenom kvalitete, a u kreiranje je bilo uključeno više od 13500 volontera. Ovaj se skup podataka često koristi za fino podešavanje jezičnih modela. Cilj je stvoriti robusne modele za dijalog otvorenog koda koji su konkurentni komercijalnim modelima.

- Ostali skupovi podataka:

- The Pile [58]: Veliki javno dostupni skup podataka (825GB) koji u sebi sadrži 22 manja, kvalitetna skupa podataka iz raznih domena kao što su znanstveni članci te drugi profesionalni izvori. Skup podataka je kompresiran pomoću Zstandarda, a sam format podataka je jsonlines [59]. Kreiran je od strane EleutherAI [60] organizacije. Ovaj skup podataka koristi se za predtreniranje jezičnih modela.
- RedPajama-Data-v2 [61]: Alternativni javno dostupni skup podataka koji sadrži veliku količinu teksta preuzetu sa CommonCrawl [62], dok se v1 više fokusirala na Wikipediju te druge izvore kao što su knjige, ArXiv, GitHub, itd, te je namijenjen za predtreniranje jezičnih modela. Tekst je filtriran kako bi se uklonili duplikati i nepotrebni podaci. Ukupno postoji preko 100 trilijuna sirovih tokena, ali nakon filtriranja i deduplikacije ostalo je 30 trilijuna tokena. Skup pokriva 5 jezika i uključuje 40+ anotiranih podataka koje se mogu koristiti za daljnju obradu i analizu. Ovaj skup podataka nije kreirala jedna osoba ili organizacija u klasičnom smislu. Umjesto toga, ovaj masivni skup podataka je rezultat rada zajednice istraživača i inženjera okupljenih oko projekta TogetherAI [63].
- Dolly 2.0 [64]: Dolly je veliki instruktivski jezični model obučen na platformi za strojno učenje Databricks. Ovaj model je licenciran za komercijalnu upotrebu. Dolly je temeljen na modelu pythia-12b i obučen je na približno 15.000 uputa/odgovora koje su generirali

zaposlenici Databricksa u različitim područjima. Najčešće se koristi za fino podešavanje velikih jezičnih modela.

#### **4.6. Zaključak**

Mehanizam pažnje značajno je transformirao strojno učenje omogućivši modelima efikasno adresiranje i analiziranje ključnih informacije iz velikih i kompleksnih skupova podataka. Njegova primjena u transformerima, posebno u obradi prirodnog jezika, donijela je izvanredne rezultate, čime je postao nezamjenjiv alat u suvremenoj praksi umjetne inteligencije. Iako su osnovni, instruktivni i chat modeli u obradi prirodnog jezika temeljeni na sličnim arhitekturama transformera, među njima postoje značajne razlike. Osnovni modeli generiraju tekst jednostavnim uzročnim modeliranjem, dok su instruktivni modeli osposobljeni za praćenje uputa, a chat modeli specijalizirani za upravljanje višekratnim dijalozima. Uz to, kvaliteta i vrsta skupa podataka igraju ključnu ulogu u treniranju ovih modela, jer direktno utječu na njihove sposobnosti i učinkovitost u stvarnim primjenama.

## 5. INŽENJERING UPITA (*engl. Prompt Engineering*)

Inženjering upita [65] (*engl. prompt engineering*) odnosi se na proces osmišljavanja i finog podešavanja upita ili uputa za jezične modele kako bi se optimizirala njihova učinkovitost i relevantnost u kontekstu zadatka.

U ovom radu bit će prikazane različite tehnike inženjeringa upita kako bi se kreirao tjedni jelovnik za osobu koja ima VLCAD [66] poremećaj. Osnovni problem pri slaganju jelovnika je količina masti koju osoba smije unijeti tijekom dana, točnije, dnevna granica koju se ne bi smjelo prekoračiti je otprilike 20 grama za dijete staro 12 godina.

Tehnike inženjeringa upita [65] koje će koristiti za pomoć pri izradi jelovnika su sljedeće:

- Inženjering upita s nula pokušaja (*engl. Zero-Shot Prompting*)
- Inženjering upita s više pokušaja (*engl. Few-Shot Prompting*)
- Inženjering upita s lančanim promišljanjem (*engl. Chain-of-Thought Prompting*)
- Generiranje uz dohvaćanje podataka (*engl. Retrieval Augmented Generation (RAG)*)
- Inženjering upita s poticanjem zaključivanja (*engl. ReAct Prompting*)

Veliki jezični model koji će se koristiti je gpt-4o [41].

### 5.1. Opće postavke velikih jezičnih modela

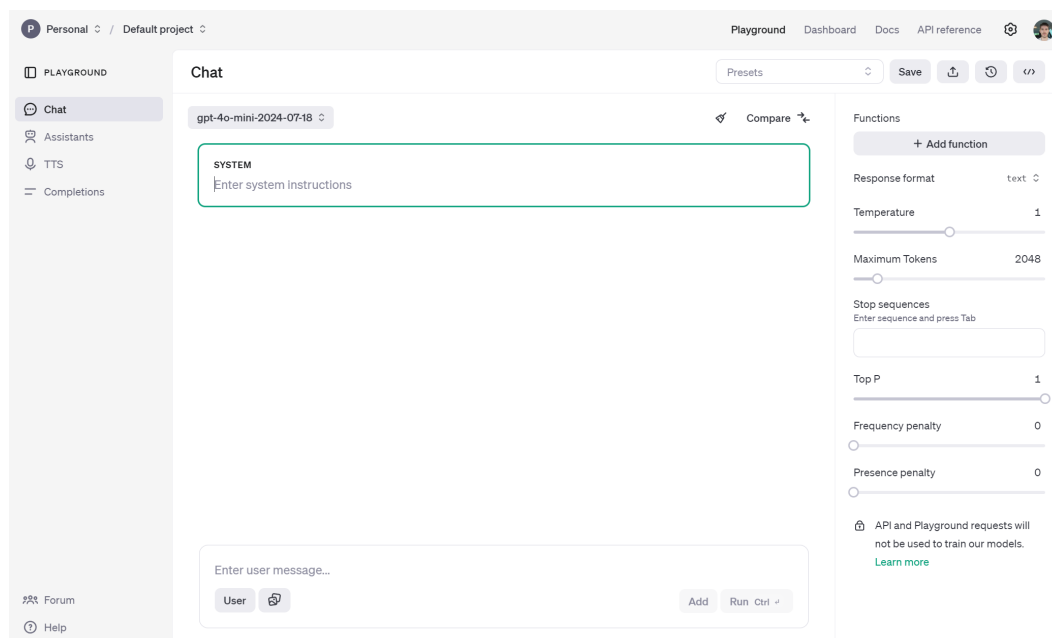
Prilikom dizajniranja i testiranja upita s velikim jezičnim modelima putem sučelja za programiranje aplikacija, obično se podešavaju različiti parametri kako bi se postigli željeni rezultati. Važno je prilagoditi ove postavke kako bi se povećala pouzdanost i kvaliteta odgovora, a ponekad je potrebno i malo eksperimentiranja kako bi se utvrdile ispravne postavke za određene slučajeve upotrebe. U nastavku su navedene uobičajene postavke koje se koriste kod različitih velikih jezičnih modela:

- **Temperatura:** Snižavanjem temperature dobivaju se deterministički rezultati, gdje se uvijek bira sljedeći najvjerojatniji token. Povećanjem temperature uvodi se više slučajnosti, što može dovesti do kreativnijih odgovora. Niža vrijednost preporučuje se za zadatke koji zahtijevaju činjenice, dok se viša vrijednost koristi za kreativne zadatke poput generiranja poezije.
- **Top P:** Ova tehnika uzorkovanja omogućuje kontrolu koliko će model biti deterministički. Niske vrijednosti osiguravaju sigurnije odgovore, dok se više vrijednosti koriste za raznolikije odgovore.

Opća preporuka [65] je mijenjati ili temperaturu ili Top P, ali ne oboje.

- **Maksimalna duljina:** Ovom postavkom se podešava broj tokena koji model generira, čime se kontrolira duljina odgovora te troškovi.
- **Zaustavni niz:** Određeni znak ili niz koji zaustavlja generiranje teksta, što omogućuje kontrolu duljine i strukture odgovora.
- **Penal učestalosti:** Primjenjuje se penal na token koji se ponavlja proporcionalno koliko se puta već pojavio, smanjujući ponavljanje riječi u odgovoru.
- **Penal prisutnosti:** Primjenjuje se penal za ponovljene tokene, ali jednako za sve ponovljene riječi. Ovo sprječava model da prečesto ponavlja fraze.

Sam izgled konzole je prikazan na slici 5.1.



Sl.5.1. Izgled ChatGPT playgrounda.

Preporučuje [65] se prilagoditi ili penal učestalosti ili penal prisutnosti, ali ne oboje.

## 5.2. Inženjering upita s nula pokušaja

Inženjering upita s nula pokušaja je tehnika u području obrade prirodnog jezika koja omogućuje jezičnim modelima generiranje odgovora na temelju upita bez potrebe za specifičnim treniranjem na određenom zadatku. Ovaj pristup koristi opće znanje modela stečeno tijekom predtreniranja na velikim količinama raznolikih podataka, omogućujući modelu da se prilagodi i odgovori na nova pitanja ili zadatke bez dodatnog prilagođavanja.

Početni upit je sljedeći:

*Kreiraj tjedni jelovnik za osobu koja ima VLCAD poremećaj u prehrani.*

Postavke jezičnog modela nisu mijenjane. Razlog tome je potreba za determinističkim tj. činjeničnim rezultatom, jer jelovnik mora sadržavati namirnice koje u sebi sadrže minimalni udio masti.

Rezultat je prikazan na slici 5.2.

The screenshot shows a chat interface with a system prompt and a user request. The assistant's response is a weekly meal plan for VLCAD deficiency, organized by day of the week. Each day includes a breakfast (Doručak), lunch (Ručak), and dinner (Večera) with specific food items and instructions. The plan is as follows:

- Ponedjeljak:**
  - Doručak: Zobena kaša sa svježim voćem (bobičasto voće) i bademovim mlijekom.
  - Užina: Šaka orašastih plodova (npr. bademi) i suhog voća (npr. grožđice).
  - Ručak: Quinoa salata s povrćem (mrkva, paprika, krastavac) i dressing na bazi maslinovog ulja.
  - Užina: Banana i jogurt s niskim udjelom masti.
  - Večera: Piletina na žaru s rižom i brokulom na pari.
- Utorak:**
  - Doručak: Smoothie od bademovog mlijeka, špinata, banane i chia sjemenki.
  - Užina: Svježi krastavci i mrkva s humusom.
  - Ručak: Sendvič s integralnim kruhom, purećom šunkom, avokadom i rajčicom.
  - Užina: Jabuka i šaka badema.
  - Večera: Bakalar sa slatkom paprikom i kvinojom.
- Srijeda:**
  - Doručak: Tost od integralnog kruha s avokadom i kuhanim jajima.
  - Užina: Smoothie od bobičastog voća i kokosovog mlijeka.
  - Ručak: Povrtna juha s lećom i integralnim rezancima.
  - Užina: Svježe grožđe i orašasti plodovi.
  - Večera: Puretina s povrćem (paprika, tikvica) na žaru i smeda riža.
- Četvrtak:**
  - Doručak: Zobena kaša sa suhim voćem (npr. marelice) i soja mlijekom.
  - Užina: Šaka badema i grickalica od mrkve.
  - Ručak: Salata s tunjevinom, integralnim tjesteninom i raznim povrćem (rikola, rajčica, masline).
  - Užina: Jogurt s niskim udjelom masti i medom.
  - Večera: Piletina s tikvicama na žaru i pire od cvjetače.
- Petak:**
  - Doručak: Chia puding s kokosovim mlijekom i svježim voćem.
  - Užina: Jabuka i suhe brusnice.
  - Ručak: Povrtna tortilla s crnim grahom, avokadom, rajčicom i kukuruzom.
  - Užina: Krastavci i humus.
  - Večera: Losos na žaru s kvinojom i špinatom na pari.
- Subota:**

At the bottom of the chat, there is a status bar: "Latency 2461ms · Finish reason: stop · 1304 tokens".

Sl.5.2. Rezultat upita (1. dio).

Slika 5.2 prikazuje rezultat upita (cijeli jelovnik se nalazi u Prilogu 1. Jelovnik: Osnovni upit), no dublja analiza jelovnika ukazuje na neke nedostatke. Određene namirnice se svakodnevno

ponavljaju (piletina) što bi trebalo izmijeniti, dok cjelovite žitarice mogu sadržavati veći udio masti nego što je preporučeno, te ih je također potrebno izbaciti iz jelovnika.

Kako bi se dobio precizniji i detaljniji jelovnik, početni upit potrebno je proširiti. Uz dodatne instrukcije o preferencijama osobe za koju se radi jelovnik, uvodi se uloga [67] nutricionista.

Uvođenjem uloge nutricionista doprinosi se poboljšanju modelovih performansi, i to na sljedeći način:

- **Specijalizacija znanja:** Dodjeljivanjem uloge nutricionista modelu se omogućuje fokusiranje na specifično područje prehrane i prehrambenih namirnica. Ovo specijalizirano znanje omogućava modelu bolje razumijevanje i interpretaciju informacija o prehrambenim potrebama, nutritivnim vrijednostima i prehrambenim smjernicama.
- **Prilagodba na specifične zahtjeve:** Kao nutricionist, model je usmjeren na pružanje preporuka koje odgovaraju specifičnim prehrambenim potrebama korisnika. To uključuje prilagođavanje jelovnika prema individualnim ciljevima, alergijama, intolerancijama i drugim posebnim zahtjevima, što rezultira preciznijim i korisnijim odgovorima.
- **Kontekstualno razumijevanje:** Uloga nutricionista pomaže modelu bolje razumjeti kontekst prehrambenih preporuka. Model može uzeti u obzir faktore kao što su zdravstveni uvjeti, ciljevi u vezi s težinom, i kulturne ili osobne prehrambene preferencije. Ovo kontekstualno razumijevanje poboljšava kvalitetu i relevantnost preporuka.
- **Korištenje specifičnih resursa:** Kao nutricionist, model može pristupiti specifičnim resursima i znanstvenim informacijama vezanim za prehranu i zdravlje. Ovo modelu omogućava pružanje savjeta i preporuka koje su u skladu s najnovijim istraživanjima i standardima u nutricionizmu.
- **Povećanje povjerenja korisnika:** Korisnici imaju tendenciju vjerovati stručnjacima s jasno definiranim ulogama. Kada model preuzima ulogu nutricionista, korisnici mogu imati veće povjerenje u točnost i korisnost preporuka, što može poboljšati iskustvo i zadovoljstvo korisnika.

Prošireni upit je sljedeći:

***Korisnik:** Pretpostavi da si vrhunski nutricionist koji ima više do 10 godina iskustva pri kreiranju jelovnika za osobe s VLCAD poremećajima u prehrani.*

*Sljedi sljedeće upute prilikom izrade jelovnika:*

- *Napravi jelovnik za 7 dana u tjednu*

- *Navedi količine te nutritivnu vrijednost za svaku namirnicu, uz broj kalorija navedi i količinu masti na 100g namirnice*
- *Ukupna količina masti u svim obrocima u danu ne bi trebala biti veća od 20 grama.*
- *Koristi namirnice koje su lako dostupne u Hrvatskoj*
- *Namirnice se mogu ponavljati svaki drugi dan*
- *Pojasni pripremu pojedinog obroka*
- *Međuobroci trebaju biti svaka 2.5 sata*
- *Uzmi u obzir da osoba voli palentu s jogurtom za doručak*
- *Uzmi u obzir da osoba voli jesti gris s Kraš ekspresom*
- *Uzmi u obzir da osoba ne voli jesti sirovo povrće*
- *Uzmi u obzir da osoba ne voli jesti sendviče*
- *Uzmi u obzir da osoba ne voli jesti kruh namazan sirom*
- *Uzmi u obzir da osoba ide u osnovnu školu, stoga 2 užine trebaju biti jednostavne i brze za pripremiti*

Rezultat proširenog upita je puno kvalitetniji jelovnik, no prilikom kreiranja istog, model je došao do limita maksimalnog broja tokena, te je odgovor prekinut, što se može vidjeti na dnu slike 5.3.

gpt-4o
Compare

- Puretina: 100 g (104 kcal, 1,5 g masti)
- Kuhana riža: 150 g (195 kcal, 0,5 g masti)
- Kuhana tikvica: 100 g (20 kcal, 0,2 g masti)
- **\*\*Priprema\*\***: Puretinu peći ili kuhati, rižu kuhati, tikvicu kuhati.
- **\*\*Nutritivna vrijednost\*\***: 319 kcal, 2,2 g masti

**\*\*Međuobrok (14:30 h)\*\***

- **\*\*Kruška\*\***: 150 g (81 kcal, 0,3 g masti)
- **\*\*Nutritivna vrijednost\*\***: 81 kcal, 0,3 g masti

**\*\*Večera (18:00 h)\*\***

- **\*\*Riblji file s povrćem\*\***
- Riblji file (oslić): 150 g (105 kcal, 0,6 g masti)
- Kuhani batat: 150 g (130 kcal, 0,2 g masti)
- Kuhana mrkva: 100 g (41 kcal, 0,1 g masti)
- **\*\*Priprema\*\***: Ribu peći ili kuhati, povrće kuhati.
- **\*\*Nutritivna vrijednost\*\***: 276 kcal, 0,9 g masti

**\*\*Ukupno dnevno\*\***: 1233 kcal, 5,7 g masti

### Četvrtak

**\*\*Doručak (7:00 h)\*\***

- **\*\*Griz s kraš ekspresom\*\***
- Griz: 50 g (182 kcal, 0,8 g masti)
- Mlijeko niskomasno: 200 ml (90 kcal, 3 g masti)
- Kraš ekspres: 5 g (20 kcal, 0,3 g masti)
- **\*\*Priprema\*\***: Griz kuhati u mlijeku dok ne postane kremast, dodati kraš ekspres.
- **\*\*Nutritivna vrijednost\*\***: 292 kcal, 4,1 g masti

**\*\*Užina (9:30 h)\*\***

- **\*\*Banana\*\***: 100 g (90 kcal, 0,3 g masti)
- **\*\*Nutritivna vrijednost\*\***: 90 kcal, 0,3 g masti

**\*\*Ručak (12:00 h)\*\***

- **\*\*Piletina s povrćem\*\***
- Pileća prsa: 100 g (165 kcal, 3,6 g masti)
- Kuhana kvinoja: 150 g (222 kcal, 3,8 g masti)
- Kuhana mrkva: 100 g (41 kcal, 0,1 g masti)
- **\*\*Priprema\*\***: Pileća prsa peći na roštilju, kuhati kvinoju i mrkvu.
- **\*\*Nutritivna vrijednost\*\***: 428 kcal, 7,5 g masti

**\*\*Međuobrok (**

ⓘ Maximum Tokens limit reached, response terminated

Latency 25148ms · Finish reason: length · 17695 tokens

Functions

+ Add function

Response format: text

Temperature: 1

Maximum Tokens: 2048

Stop sequences: Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 0

API and Playground requests will not be used to train our models. [Learn more](#)

Sl.5.3. Rezultat poboljšanog upita je došao do limita maksimalnog broja tokena.

Dubljom analizom istog može se utvrditi kako je model kreirao jelovnik prema uputama pridržavajući se preferencija osobe za koju je i kreiran. Jelovnik je kreiran za 7 dana u tjednu. Navedene su količine za sve namirnice koje se nalaze u jelovniku, kao i količina masti te broj kalorija. Dnevni unos masti ne prelazi 20g. Sve korištene namirnice su lako dostupne u Hrvatskoj. Neki obroci se ponavljaju ali ne dan za danom. Priprema pojedinih obroka je pojašnjena, te su međuobroci isplanirani za svaka 2.5 sata. Nutritivna vrijednost obroka je dobro balansirana.



### 5.3. Inženjering upita s više pokušaja

Inženjering upita s više pokušaja je tehnika u obuci jezičnih modela gdje se modelu daje nekoliko primjera (obično između 2 i 5) s uputama kako bi se poboljšala njegova sposobnost generiranja točnih odgovora za određeni zadatak. Za razliku od inženjeringa upita s nula pokušaja, gdje model mora odgovoriti na temelju općeg znanja bez ikakvih primjera, inženjering upita s više pokušaja pruža modelu male, konkretne uzorke kako bi mu pomogao bolje razumjeti kontekst i format zadatka. Ovaj pristup omogućuje modelima da se brzo prilagode i proizvedu kvalitetnije odgovore u različitim primjenama, kao što su klasifikacija teksta, prevođenje, i sažimanje, bez potrebe za dodatnim treniranjem ili velikim označenim skupovima podataka.

Postavke jezičnog modela nisu mijenjane. Razlog tome je potreba za determinističkim tj. činjeničnim rezultatom, jer jelovnik mora sadržavati namirnice koje u sebi sadrže minimalni udio masti.

#### Inženjering upita s 3 pokušaja

Sistemska poruka je sljedeća:

*Ti si vrhunski nutricionist sa više do 10 godina iskustva u kreiranju jelovnika za osobe s VLCAD poremećajem. Kreiraj dnevni jelovnik za osobu koja ima VLCAD poremećaj u prehrani u kojem dnevni unos masti nije veći od 20 grama.*

Dok je konverzacija sa modelom sljedeća:

**Korisnik:** *Primjer dnevnog jelovnika na bazi puretine:*

*(primjer dnevnog jelovnika na bazi puretine)*

**Asistent:** *Možeš li mi dati još jedan primjer na bazi piletine?*

**Korisnik:** *Primjer dnevnog jelovnika na bazi piletine:*

*(primjer dnevnog jelovnika na bazi piletine)*

**Asistent:** *Možeš li mi dati još jedan primjer na bazi teletine?*

**Korisnik:** *Primjer dnevnog jelovnika na bazi teletine:*

*(primjer dnevnog jelovnika na bazi teletine)*

**Korisnik:** *Kreiraj mi dnevni meni na bazi puretine.*

*Asistent: Naravno, evo primjera dnevnog jelovnika na bazi puretine za osobu s VLCAD poremećajem: ...*

Cijeli tok konverzacije se nalazi u Prilogu 2. Inženjering upita s 3 pokušaja, dok se dnevni jelovnik dobiven iz tog upita nalazi u Prilogu 3. Dnevni jelovnik: Inženjering upita s 3 pokušaja.

Slijed komunikacije prikazan je na slici 5.4.

The screenshot displays a chat interface with a system message, two user messages, and two system responses. The system messages provide a detailed daily diet plan for a person with VLCAD, listing meals and their nutritional content. The user messages ask for a variation of the diet plan. The interface also shows a settings panel on the right and a bottom input area.

**SYSTEM**  
Ti si vrhunski nutricionist sa više do 10 godina iskustva u kreiranju jelovnika za osobe s VLCAD poremećajem. Kreiraj dnevni jelovnik za osobu koja ima VLCAD poremećaj u prehrani u kojem dnevni unos masti nije veći od 20 grama.

**USER**  
Primjer dnevnog jelovnika na bazi puretine:

#### Doručak (7:00):  
- Griz s Kraš ekspresom  
- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)  
- 40 g griza (146 kcal, 1 g masti)  
- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

#### Međuobrok (9:30):  
- Banana (120 g, 105 kcal, 0.3 g masti)

#### Ručak (12:00):  
- Pureća prsa s povrćem  
- 150 g purećih prsa (165 kcal, 1.8 g masti)  
- 200 g kuhanog povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)

#### Međuobrok (14:30):  
- Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)

#### Večera (17:00):  
- Omlet od bjelanjaka s povrćem  
- 4 bjelanjka (68 kcal, 0.2 g masti)  
- 150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)

#### Međuobrok (19:30):  
- Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)

**ASSISTANT**  
Možeš li mi dati još jedan primjer na bazi piletine?

**USER**  
Primjer dnevnog jelovnika na bazi piletine:

#### Doručak (7:00):  
- Griz s Kraš ekspresom  
- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)  
- 40 g griza (146 kcal, 1 g masti)  
- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

#### Međuobrok (9:30):  
- Banana (120 g, 105 kcal, 0.3 g masti)

#### Ručak (12:00):  
- Piletina s rižom i povrćem  
- 100 g pilećih prsa na žaru (120 kcal, 1.5 g masti)  
- 100 g kuhane riže (130 kcal, 0.3 g masti)  
- 150 g kuhane povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)

Latency: 871ms · Finish reason: stop · 3863 tokens

SI.5.4. Prikaz inženjeringa upita s više pokušaja sa sistemskom porukom.

Analizom dobivenog jelovnika dolazi se do zaključka kako su već 3 primjera jelovnika dovoljna kako bi model mogao kreirati kvalitetan jelovnik koji zadovoljava kriterije za osobu s VLCAD poremećajem. Dnevni unos masti je unutar prihvatljive granice za osobu s VLCAD poremećajem (manje od 20 grama dnevno). Jelovnik osigurava dobar izvor proteina iz puretine, jogurta,

zobnih pahuljica i kvinoje. Jelovnik je bogat ugljikohidratima iz zobnih pahuljica, kvinoje, voća i povrća, što osigurava potrebnu energiju i vlakna za probavu. Povrće i voće osiguravaju vitamine i minerale potrebne za zdravlje. Zaključak je kako jelovnik zadovoljava uvjete niskog unosa masti i pruža uravnoteženu prehranu.

### **Inženjering upita s 5 pokušaja**

Kako bi se dobio još kvalitetniji i raznovrsniji jelovnik, modelu se šalje 5 primjera dnevnih jelovnika.

Sistemska poruka je ista kao i u inženjering upita s 3 pokušaja a modelu je poslano 5 primjera dnevnih jelovnika.

Konverzacija sa modelom je sljedeća:

**Korisnik:** *Primjer dnevnog jelovnika na bazi puretine:*

*(primjer dnevnog jelovnika na bazi puretine)*

**Asistent:** *Možeš li mi dati još jedan primjer na bazi piletine?*

**Korisnik:** *Primjer dnevnog jelovnika na bazi piletine:*

*(primjer dnevnog jelovnika na bazi piletine)*

**Asistent:** *Možeš li mi dati još jedan primjer na bazi teletine?*

**Korisnik:** *Primjer dnevnog jelovnika na bazi teletine:*

*(primjer dnevnog jelovnika na bazi teletine)*

**Asistent:** *Možeš li mi dati još jedan primjer?*

...

**Korisnik:** *Kreiraj mi dnevni meni na bazi puretine.*

**Asistent:** *Naravno, evo primjera dnevnog jelovnika na bazi puretine za osobu s VLCAD poremećajem: ...*

Cijeli tok konverzacije se nalazi u Prilogu 4. Inženjering upita s 5 pokušaja, dok se dnevni jelovnik dobiven iz tog upita nalazi u Prilogu 5. Dnevni jelovnik: Inženjering upita s 5 pokušaja.

Ovaj jelovnik pruža dovoljno energije, proteina i mikro-nutrijenata uz minimalan unos masti, što je ključno za osobe s VLCAD poremećajem. Korišteni su nisko-masni izvori proteina (kao što su

puretina i bijela riba) i izbor ugljikohidrata bogatih vlaknima koji pomažu u održavanju stabilne razine energije i sprječava nagle skokove u razini šećera u krvi. Povrće je uključeno u velikim količinama za dodatne vitamine, minerale i vlakna, dok su jogurt i rižini krekeri odabrani za zdrave međuobroke. Zaključak je kako jelovnik zadovoljava uvjete niskog unosa masti i pruža uravnoteženu prehranu.

### **Inženjering upita s 8 pokušaja**

Kako bi se dobio još kvalitetniji i raznovrsniji jelovnik, modelu se šalju 8 primjera dnevnih jelovnika.

Sistemska poruka je ista kao i u inženjeringu upita s 3 pokušaja a modelu je poslana 8 primjera dnevnih jelovnika.

Konverzacija sa modelom je sljedeća:

**Korisnik:** *Primjer dnevnog jelovnika na bazi puretine:*

*(primjer dnevnog jelovnika na bazi puretine)*

**Asistent:** *Možeš li mi dati još jedan primjer na bazi piletine?*

**Korisnik:** *Primjer dnevnog jelovnika na bazi piletine:*

*(primjer dnevnog jelovnika na bazi piletine)*

...

**Asistent:** *Možeš li mi dati još jedan primjer?*

**Korisnik:** *Primjer dnevnog jelovnika:*

*(primjer dnevnog jelovnika)*

**Korisnik:** *Kreiraj mi dnevni jelovnik.*

**Asistent:** *Naravno, evo jednog primjera dnevnog jelovnika koji je prilagođen osobama s VLCAD poremećajem, s dnevnim unosom masti ne većim od 20 grama.: ...*

Cijeli tok konverzacije se nalazi u Prilogu 6. Inženjering upita s 8 pokušaja, dok se dnevni jelovnik dobiven iz tog upita nalazi u Prilogu 7. Dnevni jelovnik: Inženjering upita s 8 pokušaja.

Jelovnik dobiven inženjeringom upita s 8 pokušaja pruža dovoljan unos kalorija te poštuje maksimalni limit unosa masti (19.4g). Plan prehrane je bogat proteinima (pileća prsa, bakalar,

kvinoja), što pomaže u održavanju mišićne mase i osjećaja sitosti. Ugljikohidrati su prisutni u obliku zobnih pahuljica, voća, i povrća, što osigurava energiju i vlakna za lakšu probavu. Voće (jabuka, naranča, borovnice) i povrće (brokula, mrkva, cvjetača) bogati su vitaminima, mineralima, antioksidansima i prehranbenim vlaknima, dok jogurt nudi kalcij i probiotike koji su važni za zdravlje kostiju i crijevne flore. Zaključak je kako jelovnik zadovoljava uvjete niskog unosa masti i pruža uravnoteženu prehranu.

Kada se usporede sva tri jelovnika, nekoliko ključnih aspekata dolaze u obzir, uključujući ravnotežu makronutrijenata [68] (proteini, ugljikohidrati i masti), raznolikost hranjivih tvari, uloge specifičnih namirnica i kako odgovaraju individualnim zdravstvenim potrebama, u ovom slučaju osobi s VLCAD poremećajem.

Analizom sva tri jelovnika s naglaskom na prilagodbu za osobu s VLCAD poremećajem, primarni fokus je bio na unosu masti ispod 20 grama dnevno, osiguravanju dovoljnog unosa proteina, ugljikohidrata, vlakana i mikronutrijenata.

Jelovnik iz inženjeringa upita s 3 pokušaja ima najviši unos vlakana koji podupire zdravu probavu. Najveći kalorijski unos, što je korisno za osobe s većim energetske potrebama. Ima izbalansirani unos proteina i ugljikohidrata, te sadrži dovoljne količine voća i povrća što čini ovaj jelovnik nutritivno bogatim.

Jelovnik iz inženjeringa upita s 5 pokušaja ima adekvatan unos proteina i vlakana uz niži unos masti. Nudi optimalni unos ugljikohidrata za održavanje energije tijekom dana, no i malo niži unos vlakana u usporedbi s ostalim jelovnicima.

Jelovnik iz inženjeringa upita s 8 pokušaja ističe se nižim udjelom masti, uz dovoljan unos proteina i vlakana. Sadrži obilje povrća, voća, te kvalitetnih izvora proteina i ugljikohidrata poput puretine i rižinih krepera. Ima visoku nutritivnu vrijednost s niskim kalorijskim unosom.

Sve tri jelovnika su prilagođena osobama s VLCAD poremećajem i zadovoljavajući su iz perspektive ukupnog dnevnog unosa masti, proteina, ugljikohidrata i vlakana. Ovisno o individualnim energetske potrebama, preferencijama okusa i specifičnih prehranbenim ciljevima, svaki od ovih jelovnika može biti prikladan.

#### **5.4. Inženjering upita s lančanim promišljanjem**

Inženjering upita s lančanim promišljanjem [69] je tehnika koja poboljšava sposobnost jezičnih modela pri rješavanju složenih zadataka korištenjem razložnog razmišljanja. Umjesto da se modelu pokaže samo jedan upit ili primjer, inženjering upita s lančanim promišljanjem pruža niz međukoraka ili obrazaca razmišljanja koji vode do konačnog odgovora. Ovaj pristup pomaže modelima da razmišljaju korak po korak, što je korisno za zadatke koji zahtijevaju logičko zaključivanje, matematičke proračune ili detaljno razumijevanje. Time se postiže bolja točnost u rješavanju problema jer model može slijediti jasan niz misli i razmotriti sve relevantne informacije prije davanja odgovora.

Postavke jezičnog modela nisu mijenjane. Razlog tome je potreba za determinističkim tj. činjeničnim rezultatom, jer jelovnik mora sadržavati namirnice koje u sebi sadrže minimalni udio masti.

Chat

Presets Save ↕ ↻ ↶

gpt-4o

Compare ↶

Functions

+ Add function

Response format text

Temperature 1

Maximum Tokens 4095

Stop sequences Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models. [Learn more](#)

SYSTEM

Ti si vrhunski nutricionist sa više do 10 godina iskustva u kreiranju jelovnika za osobe s VLCAD poremećajem. Kreiraj dnevni jelovnik za osobu koja ima VLCAD poremećaj u prehrani u kojem dnevni unos masti nije veći od 20 grama. Napiši tok misli korak po korak prilikom kreiranja jelovnika.

USER

Prilikom kreiranja jelovnika, potrebno je zadovoljiti sljedeće korake:

Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika

- Unos masti: Ograničenje unosa masti na ne više od 20 grama dnevno.
- Unos ugljikohidrata: Osoba treba dovoljno ugljikohidrata kako bi imala dovoljnu količinu energije.
- Unos proteina: Proteini su također važni za održavanje mišićne mase i funkcije tijela.
- Unos mikronutrijenata: Osigurati dovoljan unos vitamina i minerala kroz raznoliku prehranu.

Korak 2: Odabir pravilnih namirnica

- Nemasne proteinske namirnice: Pileća prsa, pureća prsa, teleći but, riba (posebno bijela riba), bjelanjci, nemasni svježi sir, leća, grah, grašak, kvinoja.
- Ugljikohidrati: Zob, kvinoja, heljda, proso, krumpir, riža, tjestenina, banane.
- Zamjene za masti: MCT ulje (ulje s masnim kiselinama srednjeg lanca) kao dodatak prehrani, s obzirom na to da se lakše metabolizira kod osoba s VLCAD.

Korak 3: Praćenje unosa masti:

- Točno evidentirati i kontrolirati unos masti.

Korak 4: Izrada jelovnika

Primjer jednog jelovnika:

### Doručak (7:30)

\*\*Zobena kaša s voćem\*\*

- 40 g zobenih pahuljica (150 kcal, 3 g masti)
- 200 ml mlijeka s niskim udjelom masti (92 kcal, 3 g masti)
- 100 g svježih borovnica (57 kcal, 0 g masti)
- \*\*Ukupno\*\*: 299 kcal, 6 g masti

### Međuobrok (10:00)

\*\*Voćna salata\*\*

- 150 g jabuka (78 kcal, 0,3 g masti)
- 100 g naranče (47 kcal, 0,2 g masti)
- \*\*Ukupno\*\*: 125 kcal, 0,5 g masti

### Ručak (12:30)

\*\*Pileći file s kvinojom i povrćem\*\*

- 100 g pilećih prsa na žaru (165 kcal, 3,6 g masti)
- 150 g kuhane kvinoje (120 kcal, 1,9 g masti)
- 200 g povrća (brokula, mrkva, cvjetača, kuhana) (85 kcal, 1 g masti)
- \*\*Ukupno\*\*: 370 kcal, 6,5 g masti

### Međuobrok (15:00)

\*\*Prirodni nezaslađeni jogurt\*\*

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)
- \*\*Ukupno\*\*: 100 kcal, 3 g masti

### Večera (18:00)

\*\*Pečeni bakalara s povrćem\*\*

- 150 g bakalara pečenog (123 kcal, 1,5 g masti)
- 200 g kuhanog povrća (špinat, brokula, šparoge) (70 kcal, 0,5 g masti)
- \*\*Ukupno\*\*: 193 kcal, 2 g masti

Latency 10801ms · Finish reason: stop · 1200 tokens

Sl.5.5. Prikaz inženjeringa upita s lančanim promišljanjem.

Kako bi model dobio uvid u tok misli prilikom kreiranja jelovnika, potrebno je detaljno razraditi svaki korak u tom toku. Ova razrada omogućava modelu jasnije razumijevanje ciljeva upita, uključujući identifikaciju nutritivnih potreba korisnika i odabir pravilnih namirnica. Kroz prikaz koraka na logičan način, model može povezati različite aspekte upita, poput ograničenja unosa masti, potrebe za energijom iz ugljikohidrata, važnosti proteina za mišićnu masu, i unos mikronutrijenata za opće zdravlje. Pružanje primjera toka misli pomaže modelu u prepoznavanju specifičnih namirnica koje ispunjavaju ove nutritivne zahtjeve, kao što su nemasne proteinske namirnice, ugljikohidrati i zamjene za masti, te u praćenju i kontroli unosa masti. Razrada toka misli modelu omogućava kontekstualizaciju i prilagodbu jelovnika prema specifičnim potrebama korisnika, što rezultira preciznijim i korisnijim jelovnikom.

Sistemska poruka je sljedeća:

*Ti si vrhunski nutricionist sa više do 10 godina iskustva u kreiranju jelovnika za osobe s VLCAD poremećajem. Kreiraj dnevni jelovnik za osobu koja ima VLCAD poremećaj u prehrani u kojem dnevni unos masti nije veći od 20 grama. Napiši tok misli korak po korak prilikom kreiranja jelovnika.*

Konverzacija sa modelom je sljedeća:

**Korisnik:** Prilikom kreiranja jelovnika, potrebno je zadovoljiti sljedeće korake:

*Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika*

- 1. Unos masti: Ograničenje unosa masti na ne više od 20 grama dnevno.*
- 2. Unos ugljikohidrata: Osoba treba dovoljno ugljikohidrata kako bi imala dovoljnu količinu energije.*
- 3. Unos proteina: Proteini su također važni za održavanje mišićne mase i funkcije tijela.*
- 4. Unos mikronutrijenata: Osigurati dovoljan unos vitamina i minerala kroz raznoliku prehranu.*

*Korak 2: Odabir pravilnih namirnica*

- 1. Nemasne proteinske namirnice: Pileća prsa, pureća prsa, teleći but, riba (posebno bijela riba), bjelanjci, nemasni svježi sir, leća, grah, grašak, kvinoja.*
- 2. Ugljikohidrati: Zob, kvinoja, heljda, proso, krumpir, riža, tjestenina, banane.*
- 3. Zamjene za masti: MCT ulje (ulje s masnim kiselinama srednjeg lanca) kao dodatak prehrani, s obzirom na to da se lakše metabolizira kod osoba s VLCAD-om.*

*Korak 3: Praćenje unosa masti:*

- Točno evidentirati i kontrolirati unos masti.*

*Korak 4: Izrada jelovnika*

*Primjer jednog jelovnika:*

*(primjer jelovnika dobiven putem inženjeringa upita s 8 pokušaja).*

**Korisnik:** Kreiraj mi dnevni jelovnik. Napiši tok misli korak po korak.

**Asistent:** ### Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika



1. **Unos masti**: Ograničiti unos masti na ne više od 20 grama dnevno.

...

### Ukupno za dan

**Ukupne kalorije**: 1284 kcal

**Ukupne masnoće**: 19.8 g masti

**Napomene**: Omlet za doručak s četiri bjelanjka pruža dobru količinu proteina bez dodatka masti, dok MCT ulje dodaje potrebne masti koje se lakše metaboliziraju za osobe s VLCAD poremećajem. Teletina služi kao glavni izvor proteina za ručak i večeru, uz nemasne priloge poput kvinoje i raznog povrća. Međuobranci su lagani, ali nutritivno bogati, što omogućava da se ukupni unos masnoće zadrži unutar preporučenih granica.

Cijeli upit sa tokom misli korak po korak nalazi se u Prilogu 8. Upit: Tok misli korak po korak, dok se cijeli jelovnik koji je model izgenerirao nalazi se u Prilogu 9. Dnevni jelovnik: Tok misli korak po korak.

Zaključak je kako je jelovnik dizajniran da zadovolji nutritivne potrebe osobe s VLCAD poremećajem, pazeći na dnevno ograničenje unosa masti te osigurava dovoljan unos proteina, ugljikohidrata, vlakana, vitamina i minerala potrebnih za zdravo tijelo. Mikronutrijenti su dobro pokriveni kroz raznoliku prehranu koja uključuje voće, povrće i mliječne proizvode. Također, unos vlakana je adekvatan, čime se podržava zdrava probava. Jedini problem u ovom jelovniku je duljina vremena potrebnog za pripremu obroka, a razlog tome je teletina koja zahtijeva dulju termičku obradu (kuhanje/pečenje) u usporedbi s drugim namirnicama poput piletine ili puretine. Kako bi se zadovoljio i taj kriterij, potreban je proširiti upit dodavanjem novog koraka za brzinu pripreme hrane.

Poboljšani upit je sljedeći:

**Korisnik**: Prilikom kreiranja jelovnika, potrebno je zadovoljiti sljedeće korake:

**Korak 1**: Identifikacija nutritivnih potreba korisnika jelovnika

...

**Korak 3**: Vrijeme pripreme obroka:

- Uzmi u obzir da osoba ide u osnovnu školu, stoga obroci trebaju biti jednostavni i brzi za pripremiti (maksimalno 60 minuta)

*Korak 4: Praćenje unosa masti:*

- *Točno evidentirati i kontrolirati unos masti.*

...

Cijeli poboljšani upit nalazi se u Prilogu 10. Poboljšani upit: Tok misli korak po korak, a rezultat poboljšanog upita nalazi se u Prilogu 11. Dnevni jelovnik: Poboljšani upit s tokom misli korak po korak.

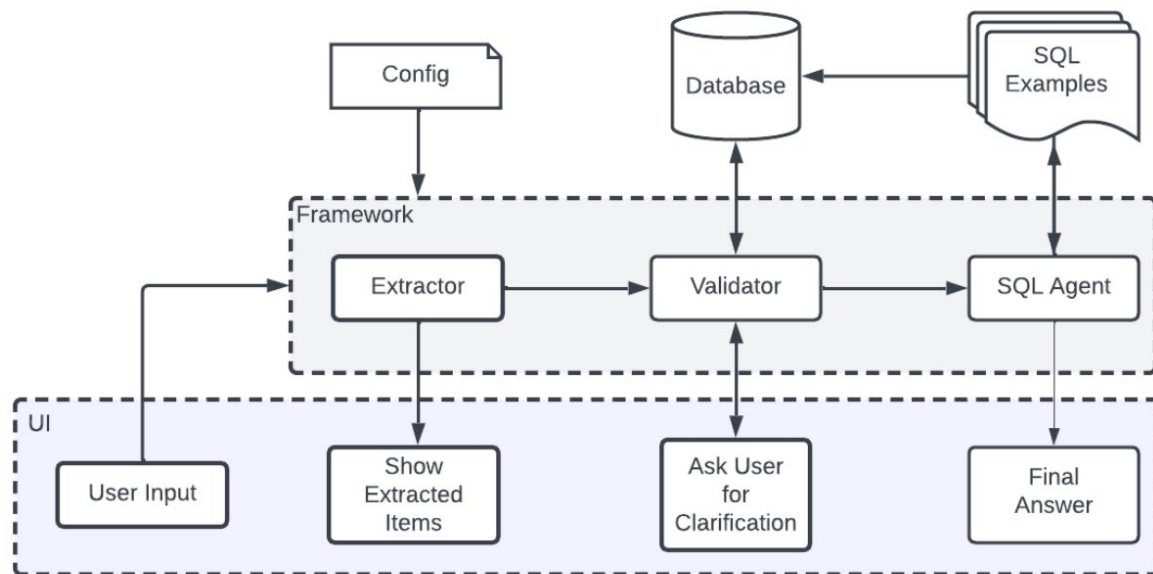
Ovaj jelovnik adekvatno zadovoljava sve makro- i mikronutritivne potrebne za dijete u osnovnoj školi. Osigurava dovoljan unos proteina za rast i razvoj, osigurava dovoljnu količinu energiju putem ugljikohidrata, zadovoljava dnevne potrebe vlakana te drži dnevni unos masti unutar preporučenog ograničenja (20 grama). Osim nutritivnih vrijednosti, jelovnik je jednostavan i brz za pripremu.

### **5.5. Generiranje uz dohvaćanje podataka (*engl. Retrieval augmented generation (RAG)*)**

Generiranje uz dohvaćanje podataka (RAG) je tehnika koja kombinira sposobnosti pretraživanja informacija i generiranja teksta unutar jezičnih modela. Ovaj pristup koristi vanjske baze podataka ili izvore znanja za dohvaćanje relevantnih informacija, koje se zatim koriste za generiranje točnih odgovora. U RAG-u, jezični model prvo pretražuje bazu podataka kako bi pronašao najrelevantnije podatke na temelju upita korisnika, a zatim generira odgovor koristeći te dohvaćene informacije. Ovaj pristup omogućuje modelima pružanje informativnijih i preciznijih odgovora, posebno za zadatke koji zahtijevaju specifične činjenice ili detalje iz velikih količina podataka, kao što su korisnička podrška, medicinska savjetovanja, ili pretraživanje istraživačkih radova.

Nažalost, tehnika generiranja uz dohvaćanje podataka ne može biti testirana zbog nedostatka vanjske baze s jelovnicima, no bit će obrađena teorijski uz primjere iz prakse, poput SoccerRAG-a [70].

SoccerRAG je radno okruženje razvijeno za upravljanje i integraciju podataka u kontekstu nogometnih statistika i analiza. Naziv "SoccerRAG" dolazi od kombinacije riječi "soccer" (nogomet) i "RAG".



Sl.5.6. Sustavni prikaz SoccerRag radnog okruženja [70].

Osnovna funkcionalnost SoccerRAG radnog okruženja prikazana je kroz sustavni tok aplikacije. Na početku, korisnik šalje upit na prirodnom jeziku, koji se potom prosljeđuje velikom jezičnom modelu zajedno s shemom svojstava i sistemskim upitom koji opisuje koja svojstva treba izdvojiti. Veliki jezični model zatim vraća popis svojstava relevantnih za upit. Sljedeće, svaka izdvojena karakteristika zatim se provjerava u odgovarajućim tablicama baze podataka koristeći algoritme za usporedbu niza kako bi se ispravile pravopisne pogreške i skraćenice. Nakon validacije, vrijednost i njezin primarni ključ dodaju se izdvojenim podacima. Pročišćeni korisnički upit integrira se sa sistemskim upitom kako bi se veliki jezični model usmjerio u izradi SQL upita koji ima za cilj rješavanje korisničkog upita. Ovi upiti zatim se izvršavaju kroz SQL lanac, koji upravlja vanjskom bazom podataka, dohvaća tražene informacije i priprema rezultate za prezentaciju korisniku. Prikaz upita prema SoccerRAG-u i njegovi odgovori mogu se vidjeti na slikama 5.7 i 5.8. SoccerRAG baza je javno dostupna na [71].

**You** 11:17 AM

List all games played by ManU in the 16-17 season.  
Give the result as a markdown table with following format  
HomeTeam AwayTeam Score Venue Attendance Date

**Extractor** 11:17 AM

Extracted properties:

```
json
[
  {
    "team_name": [
      "ManU"
    ],
    "year_season": [
      "16-17"
    ]
  }
]
```

**Cleaner** 11:17 AM

New prompt is as follows:  
List all games played by ManU in the 16-17 season.  
Give the result as a markdown table with following format  
HomeTeam AwayTeam Score Venue Attendance Date  
Updated Information:

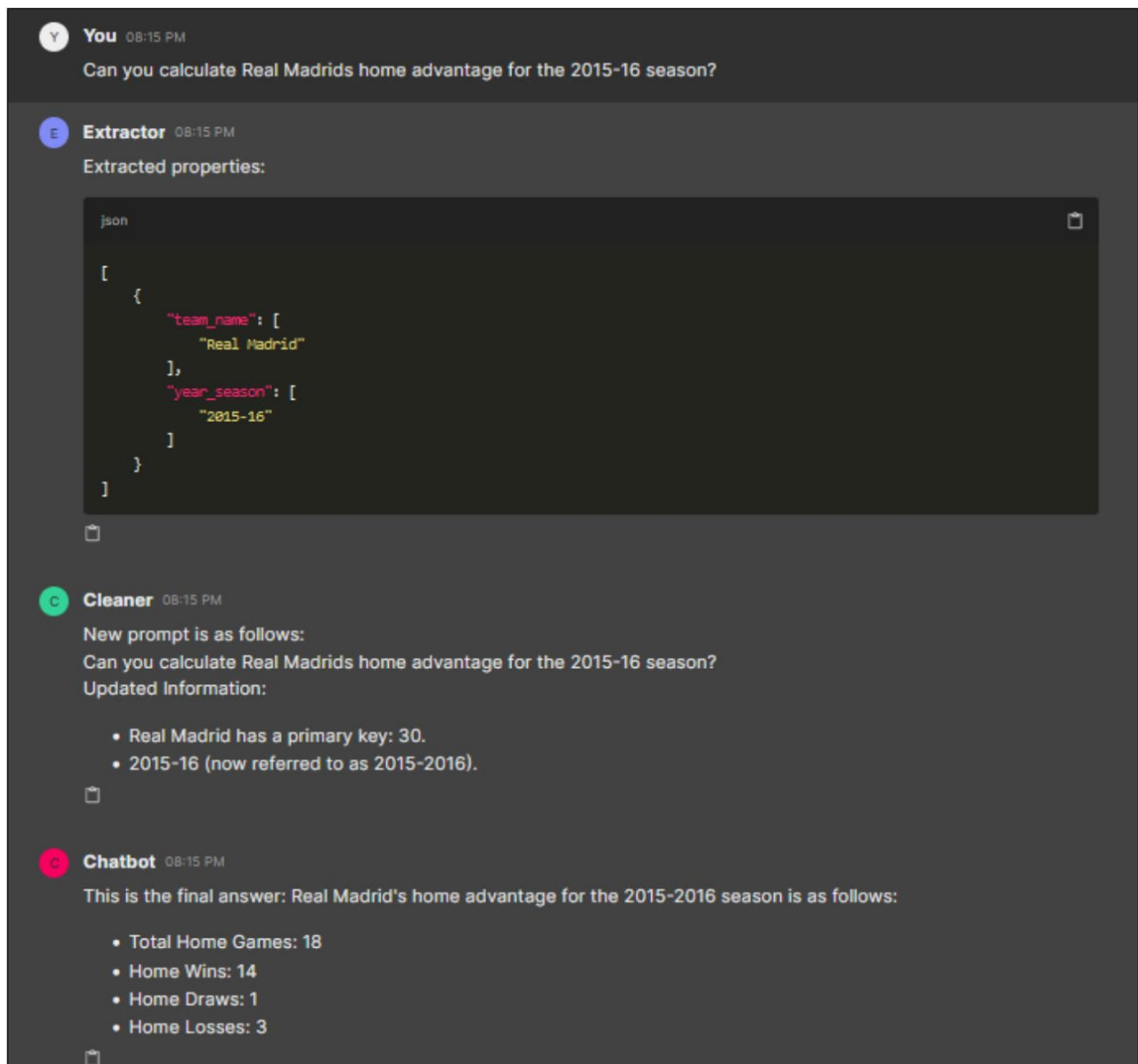
- ManU (now referred to as Manchester United) has a primary key: 7 in table teams.
- 16-17 (now referred to as 2016-2017 in table games).

**Chatbot** 11:17 AM

This is the final answer: Here are the games played by Manchester United (ManU) in the 2016-2017 season:

HomeTeam	AwayTeam	Score	Venue	Attendance	Date
Manchester United	Leicester	4 - 1	Old Trafford (Manchester)	75 256	24/09/2016 - 13:30
Liverpool	Manchester United	0 - 0	Anfield (Liverpool)	52 769	17/10/2016 - 21:00
Manchester United	Liverpool	1 - 1	Old Trafford (Manchester)	75 276	15/01/2017 - 17:00

### Sl.5.7. Prikaz RAG upita



### SI.5.8. Prikaz RAG upita 2

Neki drugi primjeri RAG-a su:

- FinanceBench [72]
- DISC-LawLLM [73]
- MedRag [74]

## 5.6. Inženjering upita s poticanjem zaključivanja (*engl. ReAct Prompting*)

Inženjering upita s poticanjem zaključivanja [75] je tehnika koja kombinira zaključivanje (*engl. reasoning*) i djelovanje (*engl. acting*) unutar jezičnih modela kako bi se poboljšala njihova sposobnost rješavanja složenih zadataka i interaktivnih scenarija. Umjesto jednostavnog generiranja odgovora na temelju statičnih upita, inženjering upita s poticanjem zaključivanja omogućuje modelu razmišljanje korak po korak dok istovremeno poduzima radnje poput postavljanja daljnjih pitanja ili dohvaćanja dodatnih informacija. Inženjering upita s poticanjem zaključivanja omogućuje modelima da budu fleksibilniji, intuitivniji i sposobni za kompleksnije zadatke razumijevanja i interakcije.

Primjer upita je sljedeći:

*Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika*

*Pojašnjenje: Jasno formulirajte specifične nutritivne potrebe osobe radi boljeg razumijevanja. Koje su osnovne nutritivne potrebe (unos masti, ugljikohidrati, proteini, vitamini i minerali)?*

*Zaključak upita: Ograničiti unos masti na ne više od 20 grama dnevno, osigurati adekvatan unos proteina i ugljikohidrata, te osigurati raznolik unos mikronutrijenata (vitamina i minerala).*

*Korak 2: Odabir pravilnih namirnica*

*Pojašnjenje: Navesti koje su namirnice prikladne za unos proteina, ugljikohidrata te kontrolu masnoće.*

- *Koje nemasne proteinske namirnice možemo koristiti?*
- *Koji izvori ugljikohidrata su najbolji za osiguranje energije?*
- *Koje zamjene za masti možemo uključiti u jelovnik?*

*Zaključak upita: Odabrane namirnice trebaju biti bogate proteinima i ugljikohidratima, a istovremeno imati nizak udio masti. Dodatak MCT ulja kao zamjene za tradicionalne masti može pomoći u lakšoj razgradnji masti kod osobe s VLCAD poremećajem.*

*Korak 3: Vrijeme pripreme obroka*

*Pojašnjenje: Pojasnite zahtjeve za vremensku učinkovitost pripreme obroka.*

- *Koliko vremena bi priprema svakog obroka trebalo trajati?*

*Zaključak upita: Obroci trebaju biti jednostavni i brzi za pripremu, maksimalno 60 minuta.*

*Korak 4: Praćenje unosa masti*

*Pojašnjenje: Kako točno pratiti i evidentirati unos masti?*

- *Koji alati ili metode se mogu koristiti za točno praćenje unosa masti?*

*Zaključak upita: Evidentirati i kontrolirati svaki unos masti unutar obroka.*

*Korak 5: Izrada jelovnika*

*Pojašnjenje: Kreiraj primjer jelovnika na temelju gore navedenih koraka.*

- *Kako strukturirati obroke tijekom dana kako bi zadovoljili sve nutritivne kriterije i vremenske zahtjeve?*

Cijeli upit nalazi se u Prilogu 12. ReART upit, dok se jelovnik nalazi u Prilogu 13. Dnevni jelovnik ReART.

Chat

Presets Save Share Refresh

gpt-4o

**SYSTEM**  
Enter system instructions

**USER**  
 Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika  
 Pojašnjenje: Jasno formulirajte specifične nutritivne potrebe osobe radi boljeg razumijevanja. Koje su osnovne nutritivne potrebe (unos masti, ugljikohidrati, proteini, vitamini i minerali)?  
 Zaključak upita: Ograničiti unos masti na ne više od 20 grama dnevno, osigurati adekvatan unos proteina i ugljikohidrata, te osigurati raznolik unos mikronutrijenata (vitamina i minerala).  
 Korak 2: Odabir pravilnih namirnica  
 Pojašnjenje: Navesti koje su namirnice prikladne za unos proteina, ugljikohidrata te kontrolu masnoće.  
 - Koje nemasne proteinske namirnice možemo koristiti?  
 - Koji izvori ugljikohidrata su najbolji za osiguranje energije?  
 - Koje zamjene za masti možemo uključiti u jelovnik?  
 Zaključak upita: Odabrane namirnice trebaju biti bogate proteinima i ugljikohidratima, a istovremeno imati nizak udio masti. Dodatak MCT ulja kao zamjene za tradicionalne masti može pomoći u laksoj razgradnji masti kod osobe s VLCAD poremećajem.  
 Korak 3: Vrijeme pripreme obroka  
 Pojašnjenje: Pojasnite zahtjeve za vremensku učinkovitost pripreme obroka.  
 - Koliko vremena bi priprema svakog obroka trebalo trajati?  
 Zaključak upita: Obroci trebaju biti jednostavni i brzi za pripremu, maksimalno 60 minuta.  
 Korak 4: Praćenje unosa masti  
 Pojašnjenje: Kako točno pratiti i evidentirati unos masti?  
 - Koji alati ili metode se mogu koristiti za točno praćenje unosa masti?  
 Zaključak upita: Evidentirati i kontrolirati svaki unos masti unutar obroka.  
 Korak 5: Izrada jelovnika  
 Pojašnjenje: Kreiraj primjer jelovnika na temelju gore navedenih koraka.  
 - Kako strukturirati obroke tijekom dana kako bi zadovoljili sve nutritivne kriterije i vremenske zahtjeve?

**USER**  
Kreiraj mi dnevni jelovnik. Pojasni tok misli korak po korak.

**ASSISTANT**  
 Da bi se izradio optimalan dnevni jelovnik po gore navedenim koracima, važno je pratiti redoslijed kako bismo osigurali sve nutritivne potrebe, odabrali prikladne namirnice, osigurali učinkovitu pripremu i kontrolirali unos masti. Evo kako bi to moglo izgledati:

### Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika  
 Nutritivne potrebe uključuju:  
 - Maksimalno 20 grama masti dnevno  
 - Adekvatni unos proteina i ugljikohidrata  
 - Raznolik unos vitamina i minerala

### Korak 2: Odabir pravilnih namirnica  
 #### Nemasne proteinske namirnice:  
 - Piletina (bez kože)  
 - Puretina (bez kože)  
 - Bijela riba (bakalar, oslić)  
 - Tofu  
 - Jaja (posebno bjelanjci)

Functions

+ Add function

Response format text

Temperature 1

Maximum Tokens 4096

Stop sequences  
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models.  
[Learn more](#)

### Sl.5.9. Prikaz inženjeringa upita s poticanjem zaključivanja.

Ovaj jelovnik pokazuje da se unos manje od 20 grama masti dnevno uspješno postiže kroz pažljiv odabir namirnica i uključivanje zdravih masnoća poput MCT ulja. Unos proteina i ugljikohidrata je adekvatan i osigurava potrebnu energiju za tijelo. Nadalje, raznolika prehrana osigurava dovoljan unos mikronutrijenata i vlakana, čime se zadovoljava većina dnevnih nutritivnih potreba. Općenito, ovakav jelovnik može biti korisno sredstvo za osobe koje trebaju kontrolirati unos masti, dok istovremeno osigurava balansirane i raznolike obroke koji podržavaju sveukupno zdravlje tijela.



## 5.7. Koncept i svrha inženjeringa upita

Koncept inženjering upita temelji se na ideji kako se jezični modeli mogu voditi prema određenim zadacima ili ciljevima pružanjem pažljivo dizajniranih upita ili uputa. Ovi upiti pomažu u usmjeravanju modela prema specifičnim vrstama odgovora ili ponašanjima koja su korisna za korisnika. Svrha inženjeringa upita je poboljšati točnost, relevantnost i koherentnost generiranih odgovora dobivenih od modela, što je osobito važno kada se modeli koriste u aplikacijama s visokim stupnjem osjetljivosti, kao što su medicinska dijagnostika, pravne analize i obrazovanje [76].

Inženjering upita također omogućuje iskorištavanje sposobnosti jezičnih modela bez potrebe za dodatnim resursima za treniranje ili fino podešavanje. Ovaj pristup je koristan jer smanjuje potrebu za dodatnim računalnim resursima i vremenskim ulaganjima koja su obično potrebna za fino podešavanje velikih jezičnih modela [16]. Na primjer, umjesto da se GPT-3 ponovno trenira za specifičan zadatak, jednostavno dizajniranje pravog upita može navesti model da pruži točne i korisne odgovore.

## 6. ZAKLJUČAK

Inženjering upita predstavlja ključnu metodologiju koja usmjerava jezične modele, kao što su GPT-4 ili Gemini, prema stvaranju kvalitetnih, relevantnih i preciznih odgovora prilagođenih specifičnim potrebama korisnika. U ovom radu prikazano je kako se različite tehnike inženjeringa upita mogu primijeniti za kreiranje jelovnika (tjednog/dnevnog) za osobu s VLCAD poremećajem. Kroz eksperimentiranje sa različitim metodama inženjeringa upita, inženjering upita s nula pokušaja, inženjering upita s više pokušaja, inženjering upita s lančanim promišljanjem te inženjeringom upita s poticanjem zaključivanja uz teorijsko sagledavanje RAG tehnike, dobiveni su jelovnici koji poštuju specifične nutritivne potrebe i ograničenja unosa masti (maksimalno 20 grama dnevno).

Analizom rezultata, utvrđeno je da je inženjering upita s više pokušaja učinkovit pristup za kreiranje raznovrsnih i nutritivno balansiranih jelovnika. Tehnika s lančanim promišljanjem pokazala se korisnom za razumijevanje koraka i logičkog toka misli, dok je inženjering upita s poticanjem zaključivanja omogućio fleksibilnost i prilagodbu jelovnika ovisno o pojedinim potrebama korisnika.

Za daljnje poboljšanje jelovnika, potrebna je integracija dodatnih baza podataka pomoću RAG-a i korištenje vanjskih resursa kako bi se osigurala još veća preciznost i informativnost odgovora. Iako su postignuti rezultati obećavajući, daljnja istraživanja i testiranja su potrebna kako bi se osigurala krajnja točnost i prilagodba na individualnoj razini.

Naposljetku, tehnike inženjeringa upita mogu značajno unaprijediti sposobnosti jezičnih modela, omogućujući im da pružaju korisne odgovore u raznim domenama, olakšavajući svakodnevne zadatke poput planiranja prehrane, ali i mnogo složenijih zadataka poput medicinske dijagnostike ili pravnih analiza.

## LITERATURA

- [1] A. Vaswani *i ostali*, „Attention Is All You Need“, 01. kolovoz 2023., *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
- [2] „OpenAI Platform“. Pristupljeno: 17. svibanj 2024. [Na internetu]. Dostupno na: <https://platform.openai.com>
- [3] „ChatGPT“. Pristupljeno: 17. svibanj 2024. [Na internetu]. Dostupno na: <https://chatgpt.com>
- [4] „Meta Llama“, Meta Llama. Pristupljeno: 17. svibanj 2024. [Na internetu]. Dostupno na: <https://llama.meta.com/>
- [5] G.-3 Demo, „Chinchilla by DeepMind | Discover AI use cases“. Pristupljeno: 17. svibanj 2024. [Na internetu]. Dostupno na: <https://gpt3demo.com/apps/chinchilla-deepmind>
- [6] „Gemini – razgovorom do novih ideja“, Gemini. Pristupljeno: 17. svibanj 2024. [Na internetu]. Dostupno na: <https://gemini.google.com>
- [7] „What Is NLP (Natural Language Processing)? | IBM“. Pristupljeno: 08. rujan 2024. [Na internetu]. Dostupno na: <https://www.ibm.com/topics/natural-language-processing>
- [8] D. Khurana, A. Koli, K. Khatter, i S. Singh, „Natural language processing: state of the art, current trends and challenges“, *Multimed. Tools Appl.*, sv. 82, izd. 3, str. 3713–3744, sij. 2023, doi: 10.1007/s11042-022-13428-4.
- [9] „How Gmail sorts your email based on your preferences“, Google Workspace Blog. Pristupljeno: 08. rujan 2024. [Na internetu]. Dostupno na: <https://workspace.google.com/blog/productivity-collaboration/how-gmail-sorts-your-email-based-on-your-preferences>
- [10] A. M. TURING, „I.—COMPUTING MACHINERY AND INTELLIGENCE“, *Mind*, sv. LIX, izd. 236, str. 433–460, lis. 1950, doi: 10.1093/mind/LIX.236.433.
- [11] „Cartesian linguistics“, *Wikipedia*. 12. veljača 2024. Pristupljeno: 17. svibanj 2024. [Na internetu]. Dostupno na: [https://en.wikipedia.org/w/index.php?title=Cartesian\\_linguistics&oldid=1206654573](https://en.wikipedia.org/w/index.php?title=Cartesian_linguistics&oldid=1206654573)
- [12] J. Devlin, M.-W. Chang, K. Lee, i K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, 24. svibanj 2019., *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
- [13] A. R. Linn, „Concise History of the Language Sciences: From the Sumerians to the Cognitivists. Ed. by E. F. K. Koerner and R. E. Asher“, *Historiogr. Linguist.*, sv. 23, izd. 1–2, str. 194–200, sij. 1996, doi: 10.1075/hl.23.1-2.23lin.
- [14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, i I. Sutskever, „Language Models are Unsupervised Multitask Learners“.
- [15] „What Are Large Language Models (LLMs)? | IBM“. Pristupljeno: 08. rujan 2024. [Na internetu]. Dostupno na: <https://www.ibm.com/topics/large-language-models>
- [16] T. B. Brown *i ostali*, „Language Models are Few-Shot Learners“, 22. srpanj 2020., *arXiv*: arXiv:2005.14165. doi: 10.48550/arXiv.2005.14165.
- [17] „GPT-4“. Pristupljeno: 08. rujan 2024. [Na internetu]. Dostupno na: <https://openai.com/index/gpt-4-research/>
- [18] „Introducing GPT-4o and more tools to ChatGPT free users“. Pristupljeno: 08. rujan 2024. [Na internetu]. Dostupno na: <https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/>
- [19] Y. Liu *i ostali*, „RoBERTa: A Robustly Optimized BERT Pretraining Approach“, 26. srpanj 2019., *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.
- [20] V. Sanh, L. Debut, J. Chaumond, i T. Wolf, „DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter“, 29. veljača 2020., *arXiv*: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.

- [21] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, i R. Soricut, „ALBERT: A Lite BERT for Self-supervised Learning of Language Representations“, 08. veljača 2020., *arXiv*: arXiv:1909.11942. doi: 10.48550/arXiv.1909.11942.
- [22] C. Raffel *i ostali*, „Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer“, 19. rujan 2023., *arXiv*: arXiv:1910.10683. doi: 10.48550/arXiv.1910.10683.
- [23] P. Sydorenko, „Top 5 applications of Large Language Models (LLMs) in legal practice“, JURDEP. Pristupljeno: 08. rujan 2024. [Na internetu]. Dostupno na: <https://medium.com/jurdep/top-5-applications-of-large-language-models-llms-in-legal-practice-d29cde9c38ef>
- [24] A. Agrawal i S. Singh, „Corpus Complexity Matters in Pretraining Language Models“, u *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustainNLP)*, N. Sadat Moosavi, I. Gurevych, Y. Hou, G. Kim, Y. J. Kim, T. Schuster, i A. Agrawal, Ur., Toronto, Canada (Hybrid): Association for Computational Linguistics, srp. 2023, str. 257–263. doi: 10.18653/v1/2023.sustainlp-1.20.
- [25] S. Shin *i ostali*, „On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model“, 08. svibanj 2022., *arXiv*: arXiv:2204.13509. Pristupljeno: 25. lipanj 2024. [Na internetu]. Dostupno na: <http://arxiv.org/abs/2204.13509>
- [26] S. Gururangan *i ostali*, „Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks“, 05. svibanj 2020., *arXiv*: arXiv:2004.10964. doi: 10.48550/arXiv.2004.10964.
- [27] „Masked language modeling“. Pristupljeno: 26. lipanj 2024. [Na internetu]. Dostupno na: [https://huggingface.co/docs/transformers/main/en/tasks/masked\\_language\\_modeling](https://huggingface.co/docs/transformers/main/en/tasks/masked_language_modeling)
- [28] F. Alfaro, M. R. Costa-jussà, i J. A. R. Fonollosa, „BERT Masked Language Modeling for Co-reference Resolution“, u *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, M. R. Costa-jussà, C. Hardmeier, W. Radford, i K. Webster, Ur., Florence, Italy: Association for Computational Linguistics, kol. 2019, str. 76–81. doi: 10.18653/v1/W19-3811.
- [29] P. Schmid, „Pre-Training BERT with Hugging Face Transformers and Habana Gaudi“. Pristupljeno: 26. lipanj 2024. [Na internetu]. Dostupno na: <https://www.philschmid.de/pre-training-bert-habana>
- [30] „Papers with Code - CoNLL 2003 (English) Benchmark (Named Entity Recognition (NER))“. Pristupljeno: 08. rujan 2024. [Na internetu]. Dostupno na: <https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003>
- [31] „Papers with Code - CoNLL 2003 Dataset“. Pristupljeno: 26. lipanj 2024. [Na internetu]. Dostupno na: <https://paperswithcode.com/dataset/conll-2003>
- [32] Y. Sun, Y. Zheng, C. Hao, i H. Qiu, „NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task--Next Sentence Prediction“, 18. listopad 2022., *arXiv*: arXiv:2109.03564. doi: 10.48550/arXiv.2109.03564.
- [33] „BERT“. Pristupljeno: 26. lipanj 2024. [Na internetu]. Dostupno na: [https://huggingface.co/docs/transformers/en/model\\_doc/bert](https://huggingface.co/docs/transformers/en/model_doc/bert)
- [34] A. Kesrwan, „Understanding Next Token Prediction: Concept To Code: 1st part!“, Medium. Pristupljeno: 26. lipanj 2024. [Na internetu]. Dostupno na: <https://medium.com/@akash.kesrwan99/understanding-next-token-prediction-concept-to-code-1st-part-7054dabda347>
- [35] „What Is Self-Supervised Learning? | IBM“. Pristupljeno: 26. lipanj 2024. [Na internetu]. Dostupno na: <https://www.ibm.com/topics/self-supervised-learning>
- [36] „How Transformers Work: A Detailed Exploration of Transformer Architecture“. Pristupljeno: 17. svibanj 2024. [Na internetu]. Dostupno na: <https://www.datacamp.com/tutorial/how-transformers-work>
- [37] „What are Recurrent Neural Networks? | IBM“. Pristupljeno: 02. lipanj 2024. [Na internetu]. Dostupno na: <https://www.ibm.com/topics/recurrent-neural-networks>

- [38] „Transformer Architecture: The Positional Encoding - Amirhossein Kazemnejad’s Blog“. Pristupljeno: 02. lipanj 2024. [Na internetu]. Dostupno na: [https://kazemnejad.com/blog/transformer\\_architecture\\_positional\\_encoding/](https://kazemnejad.com/blog/transformer_architecture_positional_encoding/)
- [39] J. Brownlee, „A Gentle Introduction to the Rectified Linear Unit (ReLU)“, *MachineLearningMastery.com*. Pristupljeno: 27. lipanj 2024. [Na internetu]. Dostupno na: <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/>
- [40] Y. Wu *i ostali*, „Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation“, 08. listopad 2016., *arXiv*: arXiv:1609.08144. doi: 10.48550/arXiv.1609.08144.
- [41] „Hello GPT-4o“. Pristupljeno: 16. kolovoz 2024. [Na internetu]. Dostupno na: <https://openai.com/index/hello-gpt-4o/>
- [42] FunCry, „In Depth Understanding of Attention Mechanism (Part II) - Scaled Dot-Product Attention and Example“, *Medium*. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: <https://medium.com/@funcry/in-depth-understanding-of-attention-mechanism-part-ii-scaled-dot-product-attention-and-its-7743804e610e>
- [43] S. R. PhD, „Understanding and Coding Self-Attention, Multi-Head Attention, Cross-Attention, and Causal-Attention in LLMs“. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: <https://magazine.sebastianraschka.com/p/understanding-and-coding-self-attention>
- [44] M. Gheini, X. Ren, i J. May, „Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation“, u *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, i S. W. Yih, Ur., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, stu. 2021, str. 1754–1765. doi: 10.18653/v1/2021.emnlp-main.132.
- [45] I. Beltagy, M. E. Peters, i A. Cohan, „Longformer: The Long-Document Transformer“, 02. prosinac 2020., *arXiv*: arXiv:2004.05150. doi: 10.48550/arXiv.2004.05150.
- [46] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, i S. Sanghai, „GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints“, 23. prosinac 2023., *arXiv*: arXiv:2305.13245. doi: 10.48550/arXiv.2305.13245.
- [47] F. Chiusano, „Two minutes NLP — Visualizing Global vs Local Attention“, *NLPlanet*. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: <https://medium.com/nlplanet/two-minutes-nlp-visualizing-global-vs-local-attention-c61b42758019>
- [48] M. Zaheer *i ostali*, „Big Bird: Transformers for Longer Sequences“, 08. siječanj 2021., *arXiv*: arXiv:2007.14062. doi: 10.48550/arXiv.2007.14062.
- [49] G. Yenduri *i ostali*, „Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions“, 21. svibanj 2023., *arXiv*: arXiv:2305.10435. doi: 10.48550/arXiv.2305.10435.
- [50] „Causal language modeling“. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: [https://huggingface.co/docs/transformers/tasks/language\\_modeling](https://huggingface.co/docs/transformers/tasks/language_modeling)
- [51] T. Vykruta, „Understanding Causal LLM’s, Masked LLM’s, and Seq2Seq: A Guide to Language Model Training...“, *Medium*. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: [https://medium.com/@tom\\_21755/understanding-causal-llms-masked-llm-s-and-seq2seq-a-guide-to-language-model-training-d4457bbd07fa](https://medium.com/@tom_21755/understanding-causal-llms-masked-llm-s-and-seq2seq-a-guide-to-language-model-training-d4457bbd07fa)
- [52] L. Ouyang *i ostali*, „Training language models to follow instructions with human feedback“.
- [53] „What is an instruct model? - Instruction and Chat Fine-Tuning“, *Community.aws*. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: <https://community.aws/content/2ZVa61RxToXUFzcuY8Hbut6L150/what-is-an-instruct-model>

- [54] „Illustrating Reinforcement Learning from Human Feedback (RLHF)“. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: <https://huggingface.co/blog/rlhf>
- [55] „Chat Models | LangChain“. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: [https://python.langchain.com/v0.1/docs/modules/model\\_io/chat/](https://python.langchain.com/v0.1/docs/modules/model_io/chat/)
- [56] „Stanford CRFM“. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: <https://crfm.stanford.edu/2023/03/13/alpaca.html>
- [57] „OA\_Paper\_2023\_04\_15.pdf“, Google Docs. Pristupljeno: 10. rujan 2024. [Na internetu]. Dostupno na: [https://drive.google.com/file/d/10iR5hKwFqAKhL3umx8muOWSRm7hs5FqX/view?usp=embed\\_facebook](https://drive.google.com/file/d/10iR5hKwFqAKhL3umx8muOWSRm7hs5FqX/view?usp=embed_facebook)
- [58] „The Pile“. Pristupljeno: 04. rujan 2024. [Na internetu]. Dostupno na: <https://pile.eleuther.ai/>
- [59] „JSON Lines“. Pristupljeno: 10. rujan 2024. [Na internetu]. Dostupno na: <https://jsonlines.org/>
- [60] „EleutherAI“, EleutherAI. Pristupljeno: 10. rujan 2024. [Na internetu]. Dostupno na: <https://www.eleuther.ai>
- [61] „RedPajama-Data-v2: An open dataset with 30 trillion tokens for training large language models“. Pristupljeno: 10. rujan 2024. [Na internetu]. Dostupno na: <https://www.together.ai/blog/redpajama-data-v2>
- [62] „Common Crawl - Open Repository of Web Crawl Data“. Pristupljeno: 10. rujan 2024. [Na internetu]. Dostupno na: <https://commoncrawl.org/>
- [63] „Together AI“. Pristupljeno: 10. rujan 2024. [Na internetu]. Dostupno na: <https://www.together.ai/>
- [64] *databrickslabs/dolly*. (10. rujan 2024.). Python. Databricks Labs. Pristupljeno: 10. rujan 2024. [Na internetu]. Dostupno na: <https://github.com/databrickslabs/dolly>
- [65] „Prompt Engineering Guide“. Pristupljeno: 11. rujan 2024. [Na internetu]. Dostupno na: <https://www.promptingguide.ai/techniques>
- [66] „Very long-chain acyl-CoA dehydrogenase deficiency: MedlinePlus Genetics“. Pristupljeno: 11. rujan 2024. [Na internetu]. Dostupno na: <https://medlineplus.gov/genetics/condition/very-long-chain-acyl-coa-dehydrogenase-deficiency/>
- [67] Y. Shao, L. Li, J. Dai, i X. Qiu, „Character-LLM: A Trainable Agent for Role-Playing“, u *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, str. 13153–13187. doi: 10.18653/v1/2023.emnlp-main.814.
- [68] H. Y. Life, „Makronutrijenti i mikronutrijenti u prehrani“, Hug Your Life. Pristupljeno: 15. rujan 2024. [Na internetu]. Dostupno na: <https://hugyourlife.hr/blog/makro-i-mikro-nutrijenti/>
- [69] J. Wei *i ostali*, „Chain-of-Thought Prompting Elicits Reasoning in Large Language Models“, 10. siječanj 2023., *arXiv*: arXiv:2201.11903. doi: 10.48550/arXiv.2201.11903.
- [70] A. T. Strand, S. Gautam, C. Midoglu, i P. Halvorsen, „Demo: Soccer Information Retrieval via Natural Queries using SoccerRAG“, 22. srpanj 2024., *arXiv*: arXiv:2406.01280. doi: 10.48550/arXiv.2406.01280.
- [71] *simula/soccer-rag*. (16. srpanj 2024.). Python. Simula. Pristupljeno: 15. rujan 2024. [Na internetu]. Dostupno na: <https://github.com/simula/soccer-rag>
- [72] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, i B. Vidgen, „FinanceBench: A New Benchmark for Financial Question Answering“, 20. studeni 2023., *arXiv*: arXiv:2311.11944. doi: 10.48550/arXiv.2311.11944.
- [73] S. Yue *i ostali*, „DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services“, 23. rujan 2023., *arXiv*: arXiv:2309.11325. doi: 10.48550/arXiv.2309.11325.

- [74] „Benchmarking Retrieval-Augmented Generation for Medicine“. Pristupljeno: 15. rujan 2024. [Na internetu]. Dostupno na: <https://arxiv.org/html/2402.13178v2>
- [75] S. Yao i ostali, „ReAct: Synergizing Reasoning and Acting in Language Models“, 09. ožujak 2023., *arXiv*: arXiv:2210.03629. doi: 10.48550/arXiv.2210.03629.
- [76] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, i G. Neubig, „Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing“, arXiv.org. Pristupljeno: 05. rujan 2024. [Na internetu]. Dostupno na: <https://arxiv.org/abs/2107.13586v1>

## SAŽETAK

Ovaj rad bavi se primjenom i razvojem tehnika obrade prirodnog jezika kroz upotrebu velikih jezičnih modela poput GPT-a, BERT-a i drugih naprednih modela temeljenih na arhitekturi transformera. Glavni problem koji rad obrađuje je kriranje jelovnika za osobu s VLCAD poremećajem. Kao smjernica za rješavanje ovog problema uveden je koncept inženjeringa upita, koji omogućava korisnicima da kreiraju specifične upite kako bi precizno usmjerili modele prema željenim izlazima. U radu su opisane različite metode i tehnike inženjeringa upita, uključujući dizajniranje i prilagodbu uputa modelu, te je istražena njihova učinkovitost u kontroliranju rezultata. Također su analizirani napredni modeli, poput GPT-4 i BERT-ai. Rezultat rada pokazuje da pažljivo oblikovani upiti mogu značajno poboljšati kvalitetu generiranih odgovora, smanjiti pristranost te omogućiti bolje korisničko iskustvo u interakcijama s AI sustavima.

Ključne riječi: inženjering upita, prirodna obrada jezika, samo-pažnja, transformer, veliki jezični model.



## **ABSTRACT**

### **Prompt Engineering for Large Language Models**

This paper focuses on the application and development of natural language processing techniques through the use of large language models such as GPT, BERT, and other advanced models based on transformer architecture. The main issue addressed in the paper is the creation of a menu for a person with VLCAD disorder. As a guideline for solving this problem, the concept of prompt engineering was introduced, allowing users to create specific prompts to precisely guide the models toward desired outputs. The paper describes various methods and techniques of prompt engineering, including the design and adaptation of instructions for the model, and explores their effectiveness in controlling the results. Advanced models, such as GPT-4 and BERT, are also analyzed. The findings show that carefully crafted prompts can significantly improve the quality of generated responses, reduce bias, and enhance the user experience in interactions with AI systems.

Keywords: Large language models, natural language processing, prompt engineering, self-attention, transformer.

## **ŽIVOTOPIS**

Ivan Martinović rođen je 16. veljače 1990. u Slavonskom Brodu. Pohađao je osnovnu školu OŠ Jagode Truhelke u Osijeku, nakon koje upisuje III. Gimnaziju u Osijeku. Nakon završetka srednje škole, upisuje preddiplomski sveučilišni studij računarstva na Fakultetu elektrotehnike, računarstva i informacijskih tehnologija Osijek. Po završetku preddiplomskog studija upisuje diplomski sveučilišni studij računarstva, izborni blok Procesno računarstvo.

## PRILOZI

### Prilog 1. Jelovnik: Osnovni upit

*VLCAD (Very Long Chain Acyl-CoA Dehydrogenase) poremećaj je genetski uvjetovana metabolička bolest koja utječe na sposobnost tijela da razgrađuje vrlo duge lance masnih kiselina. Osobe s ovim poremećajem trebaju prilagoditi prehranu kako bi izbjegle životinjske masti, uključile dovoljno ugljikohidrata i konzumirale srednjelančane trigliceride (MCT). U suradnji s nutricionistom ili liječnikom, evo primjer tjednog jelovnika za osobu s VLCAD:*

#### ### Ponedjeljak

- **Doručak**: Zobena kaša sa svježim voćem (bobičasto voće) i bademovim mlijekom.
- **Užina**: Šaka orašastih plodova (npr. bademi) i suhog voća (npr. groždice).
- **Ručak**: Quinoa salata s povrćem (mrkva, paprika, krastavac) i dressing na bazi maslinovog ulja.
- **Užina**: Banana i jogurt s niskim udjelom masti.
- **Večera**: Piletina na žaru s rižom i brokulom na pari.

#### ### Utorak

- **Doručak**: Smoothie od bademovog mlijeka, špinata, banane i chia sjemenki.
- **Užina**: Svježi krastavci i mrkva s humusom.
- **Ručak**: Sendvič s integralnim kruhom, purećom šunkom, avokadom i rajčicom.
- **Užina**: Jabuka i šaka badema.
- **Večera**: Bakalar sa slatkom paprikom i kvinojom.

#### ### Srijeda

- **Doručak**: Tost od integralnog kruha s avokadom i kuhanim jajima.
- **Užina**: Smoothie od bobičastog voća i kokosovog mlijeka.
- **Ručak**: Povrtna juha s lećom i integralnim rezancima.
- **Užina**: Svježe grožđe i orašasti plodovi.
- **Večera**: Puretina s povrćem (paprika, tikvica) na žaru i smeđa riža.

#### #### Četvrtak

- **Doručak**: Zobena kaša sa suhim voćem (npr. marelice) i soja mlijekom.
- **Užina**: Šaka badema i grickalica od mrkve.
- **Ručak**: Salata s tunjevinom, integralnim tjesteninom i raznim povrćem (rikola, rajčica, masline).
- **Užina**: Jogurt s niskim udjelom masti i medom.
- **Večera**: Piletina s tikvicama na žaru i pire od cvjetače.

#### #### Petak

- **Doručak**: Chia puding s kokosovim mlijekom i svježim voćem.
- **Užina**: Jabuka i suhe brusnice.
- **Ručak**: Povrtna tortilla s crnim grahom, avokadom, rajčicom i kukuruzom.
- **Užina**: Krastavci i humus.
- **Večera**: Losos na žaru s kvinojom i špinatom na pari.

#### #### Subota

- **Doručak**: Smoothie od bademovog mlijeka, banana i špinata.
- **Užina**: Mrkva i paprika s humusom.
- **Ručak**: Tjestenina od integralnog brašna s povrćem (šampinjoni, paprike) i rajčicama.
- **Užina**: Jogurt s niskim udjelom masti i medom.
- **Večera**: Pileća prsa s tikvicama na grilu i prokulicama na pari.

#### #### Nedjelja

- **Doručak**: Tost od integralnog kruha s avokadom i svježe rajčice.
- **Užina**: Jabuka i šaka orašastih plodova.
- **Ručak**: Pečeni bakalar s povrćem (brokula, mrkva) i integralnom rižom.
- **Užina**: Smoothie od bobičastog voća i kokosovog mlijeka.

- **Večera**: Povrtni stir-fry (paprika, brokula, mrkva, još povrća po želji) s piletinom i smeđom rižom.

### Savjeti:

1. **Izbjegavajte visoko masne životinjske proizvode**: Pokušajte izbjegavati proizvode s visokim udjelom masnoće.
2. **Koristite MCT ulje**: Srednjelančane trigliceride (MCT) možete dobiti iz MCT ulja koje dodajete u smoothieje, kaše i druge obroke.
3. **Održavajte redovite obroke i užine**: Stabilnost šećera u krvi je ključna.
4. **Hidratacija**: Pijte dosta vode tijekom dana.

Uvijek je najbolje konzultirati se s nutricionistom ili liječnikom kako bi se individualne potrebe pravilno zadovoljile.

## **Prilog 2. Inženjering upita s 3 pokušaja**

Sistemska poruka:

Ti si vrhunski nutricionist sa više do 10 godina iskustva u kreiranju jelovnika za osobe s VLCAD poremećajem. Kreiraj dnevni jelovnik za osobu koja ima VLCAD poremećaj u prehrani u kojem dnevni unos masti nije veći od 20 grama. Napiši tok misli korak po korak prilikom kreiranja jelovnika.

Korisnik: Primjer dnevnog jelovnika na bazi puretine:

##### Doručak (7:00):

- Griz s Kraš ekspresom

- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)

- 40 g griza (146 kcal, 1 g masti)

- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

##### Međubrok (9:30):

- Banana (120 g, 105 kcal, 0.3 g masti)

##### Ručak (12:00):

- *Pureća prsa s povrćem*

- 150 g purećih prsa (165 kcal, 1.8 g masti)

- 200 g kuhanog povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)

##### *Međuobrok (14:30):*

- *Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)*

##### *Večera (17:00):*

- *Omlet od bjelanjaka s povrćem*

- 4 bjelanjaka (68 kcal, 0.2 g masti)

- 150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)

##### *Međuobrok (19:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

*Asistent: Možeš li mi dati još jedan primjer na bazi piletine?*

*Korisnik: Primjer dnevnog jelovnika na bazi piletine:*

##### *Doručak (7:00):*

- *Griz s Kraš ekspresom*

- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)

- 40 g griza (146 kcal, 1 g masti)

- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

##### *Međuobrok (9:30):*

- *Banana (120 g, 105 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Piletina s rižom i povrćem*

- 100 g pilećih prsa na žaru (120 kcal, 1.5 g masti)

- 100 g kuhane riže (130 kcal, 0.3 g masti)

- 150 g kuhanog povrća (mrkva, grašak) (70 kcal, 0.4 g masti)

##### Međubrok (14:30):

- Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)

##### Večera (17:00):

- Omlet od bjelanjaka s povrćem

- 4 bjelanjka (68 kcal, 0.2 g masti)

- 150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)

##### Međubrok (19:30):

- Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)

...

Asistent: Možeš li mi dati još jedan primjer?

Korisnik: Primjer dnevnog jelovnika na bazi teletine:

##### Doručak (7:00):

- Palenta s jogurtom

- 50 g palente (167 kcal, 0.5 g masti)

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)

##### Međubrok (9:30):

- Jabuka (150 g, 78 kcal, 0.3 g masti)

##### Ručak (12:00):

- Kuhana teletina s povrćem

- 150 g telećeg buta (210 kcal, 8 g masti)

- 200 g kuhanog povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)

##### Međubrok (14:30):

- Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)

##### Večera (17:00):

- Omlet od bjelanjaka s povrćem

- 4 bjelanjka (68 kcal, 0.2 g masti)

- 150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)

##### Međubrok (19:30):

- Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)

(primjer dnevnog jelovnika)

Korisnik: Kreiraj mi dnevni jelovnik.

### **Prilog 3. Dnevni jelovnik: Inženjering upita s 3 pokušaja**

Naravno, evo primjer dnevnog jelovnika na bazi puretine s ukupnim unosom masti koji ne prelazi 20 grama:

##### Doručak (7:00):

- Zobene pahuljice s voćem

- 40 g zobelih pahuljica (150 kcal, 3 g masti)

- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)

- 100 g jagoda (32 kcal, 0.3 g masti)

##### Međubrok (9:30):

- Banana (120 g, 105 kcal, 0.3 g masti)

##### Ručak (12:00):

- Pileća prsa s kvinojom i povrćem

- 150 g purećih prsa na žaru (165 kcal, 1.8 g masti)

- 100 g kuhane kvinoje (120 kcal, 2 g masti)

- 200 g kuhanog povrća (mrkva, brokula, paprike) (85 kcal, 1 g masti)

##### Međubrok (14:30):

- Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)



##### Večera (17:00):

- Pureće okruglice s povrtnom salatom

- 100 g purećih okruglica (110 kcal, 2 g masti)

- 200 g salate (miješano zelje, krastavci, rajčica) s malo limunovog soka (60 kcal, 0.5 g masti)

##### Međubrok (19:30):

- Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)

### Ukupno:

- Energetski unos: ~1202 kcal

- Ukupni unos masti: ~19.3 g masti

Ovako složen jelovnik prilagođen je VLCAD poremećaju, s ciljem da se održi niski unos masti dok se osigurava adekvatan unos kalorija i hranjivih tvari.

Prilog 3. Dnevni jelovnik: Inženjering upita s 5 pokušaja

Naravno! Evo dnevnog jelovnika na bazi puretine, uz ograničenje da dnevni unos masti ne prelazi 20 grama:

### Doručak (7:00):

- \*\*Zobena kaša s bobičastim voćem\*\*

- 40 g zobenih pahuljica (150 kcal, 2.5 g masti)

- 200 ml mlijeka s niskim udjelom masti (1.5% mm) (92 kcal, 3 g masti)

- 50 g bobičastog voća (25 kcal, 0.2 g masti)

##### Međubrok (9:30):

- \*\*Jabuka\*\*

- 150 g jabuka (78 kcal, 0.3 g masti)

### Ručak (12:00):

- \*\*Puretina s kvinojom i povrćem\*\*

- 150 g purećih prsa na žaru (165 kcal, 1.8 g masti)

- 100 g kuhane kvinoje (120 kcal, 1.9 g masti)
- 200 g kuhanog povrća (brokula, mrkva, tikvica) (85 kcal, 1 g masti)

#### Međubrok (14:30):

- **\*\*Rižin krekeri s niskomasnim sirom\*\***
- 20 g rižinih krekeri (76 kcal, 0.4 g masti)
- 50 g niskomasnog sira (40 kcal, 1 g masti)

#### Večera (17:00):

- **\*\*Minestrone juha s puretinom\*\***
- 100 g purećih prsa na kockice (110 kcal, 1.2 g masti)
- 200 g povrća (mrkva, luk, celer, rajčica) u juhi (70 kcal, 0.5 g masti)
- 150 ml povrtnog temeljca (15 kcal, 0 g masti)

#### Međubrok (19:30):

- **\*\*Jogurt s niskim udjelom masti\*\***
- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)

#### Ukupni dnevni unos masti: 16.8 grama

*Nadam se da je ovaj jelovnik u skladu s tvojim potrebama. Ako imaš još neke specifične preferencije ili potrebe, slobodno ih navedi!*

#### **Priloga 4. Inženjering upita s 5 pokušaja**

*Sistemska poruka:*

*Ti si vrhunski nutricionist sa više do 10 godina iskustva u kreiranju jelovnika za osobe s VLCAD poremećajem. Kreiraj dnevni jelovnik za osobu koja ima VLCAD poremećaj u prehrani u kojem dnevni unos masti nije veći od 20 grama. Napiši tok misli korak po korak prilikom kreiranja jelovnika.*

*Korisnik: Primjer dnevnog jelovnika na bazi puretine:*

##### Doručak (7:00):

- Griz s Kraš ekspresom

- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)

- 40 g griza (146 kcal, 1 g masti)

- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

##### Međubrok (9:30):

- Banana (120 g, 105 kcal, 0.3 g masti)

##### Ručak (12:00):

- Pureća prsa s povrćem

- 150 g purećih prsa (165 kcal, 1.8 g masti)

- 200 g kuhanog povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)

##### Međubrok (14:30):

- Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)

##### Večera (17:00):

- Omlet od bjelanjaka s povrćem

- 4 bjelanjka (68 kcal, 0.2 g masti)

- 150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)

##### Međubrok (19:30):

- Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)

Asistent: Možeš li mi dati još jedan primjer na bazi piletine?

Korisnik: Primjer dnevnog jelovnika na bazi piletine:

##### Doručak (7:00):

- Griz s Kraš ekspresom

- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)

- 40 g griza (146 kcal, 1 g masti)

- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

##### *Međuobrok (9:30):*

- *Banana (120 g, 105 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Piletina s rižom i povrćem*

- *100 g pilećih prsa na žaru (120 kcal, 1.5 g masti)*

- *100 g kuhane riže (130 kcal, 0.3 g masti)*

- *150 g kuhanog povrća (mrkva, grašak) (70 kcal, 0.4 g masti)*

##### *Međuobrok (14:30):*

- *Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)*

##### *Večera (17:00):*

- *Omlet od bjelanjaka s povrćem*

- *4 bjelanjaka (68 kcal, 0.2 g masti)*

- *150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)*

##### *Međuobrok (19:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

...

*Asistent: Možeš li mi dati još jedan primjer na bazi teletine?*

*Korisnik: Primjer dnevnog jelovnika na bazi teletine:*

##### *Doručak (7:00):*

- *Palenta s jogurtom*

- *50 g palente (167 kcal, 0.5 g masti)*

- *200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)*

##### *Međuobrok (9:30):*

- *Jabuka (150 g, 78 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Kuhana teletina s povrćem*

- *150 g telećeg buta (210 kcal, 8 g masti)*

- *200 g kuhanog povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)*

##### *Međuobrok (14:30):*

- *Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)*

##### *Večera (17:00):*

- *Omlet od bjelanjaka s povrćem*

- *4 bjelanjka (68 kcal, 0.2 g masti)*

- *150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)*

##### *Međuobrok (19:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

*Asistent: Možeš li mi dati još jedan primjer?*

*Korisnik: Primjer dnevnog jelovnika na bazi ribe:*

##### *Doručak (7:00):*

- *Palenta s jogurtom*

- *50 g palente (167 kcal, 0.5 g masti)*

- *200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)*

##### *Međuobrok (9:30):*

- *Jabuka (150 g, 78 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Pečena riba s povrćem*

- *100 g pečene bijele ribe (82 kcal, 1 g masti)*

- *200 g kuhanog povrća (brokula, mrkva, cvjetača) (100 kcal, 1 g masti)*

##### *Međuobrok (14:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

##### *Večera (17:00):*

- *Pečeni bakalar s povrćem*

- *100 g bakalara (82 kcal, 1 g masti)*

- *200 g kuhanog povrća (brokula, mrkva, cvjetača) (100 kcal, 1 g masti)*

##### *Međuobrok (19:30):*

- *Mandarina (140 g, 65 kcal, 0.2 g masti)*

*Asistent: Možeš li mi dati još jedan primjer?*

*Korisnik: Primjer dnevnog jelovnika:*

##### *Doručak (7:00):*

- *Palenta s jogurtom*

- *50 g palente (167 kcal, 0.5 g masti)*

- *200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)*

##### *Međuobrok (9:30):*

- *Jabuka (150 g, 78 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Pečena piletina s pireom od krumpira*

- *100 g pečene piletine (165 kcal, 3.6 g masti)*

- *150 g pire krumpira s maslinovim uljem (120 kcal, 3 g masti)*

##### *Međuobrok (14:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

##### *Večera (17:00):*

- *Pečena riba s povrćem*

- 100 g pečene bijele ribe (82 kcal, 1 g masti)
- 200 g kuhanog povrća (brokula, mrkva, cvjetača) (100 kcal, 1 g masti)

#### Međubrok (19:30):

- Mandarina (140 g, 65 kcal, 0.2 g masti)

Korisnik: Kreiraj mi dnevni meni na bazi puretine.

### **Prilog 5. Dnevni jelovnik: Inženjering upita s 5 pokušaja**

Naravno! Evo dnevnog jelovnika na bazi puretine, uz ograničenje da dnevni unos masti ne prelazi 20 grama:

### Doručak (7:00):

- **\*\*Zobena kaša s bobičastim voćem\*\***
  - 40 g zobenih pahuljica (150 kcal, 2.5 g masti)
  - 200 ml mlijeka s niskim udjelom masti (1.5% mm) (92 kcal, 3 g masti)
  - 50 g bobičastog voća (25 kcal, 0.2 g masti)

### Međubrok (9:30):

- **\*\*Jabuka\*\***
  - 150 g jabuka (78 kcal, 0.3 g masti)

### Ručak (12:00):

- **\*\*Puretina s kvinojom i povrćem\*\***
  - 150 g purećih prsa na žaru (165 kcal, 1.8 g masti)
  - 100 g kuhane kvinoje (120 kcal, 1.9 g masti)
  - 200 g kuhanog povrća (brokula, mrkva, tikvica) (85 kcal, 1 g masti)

### Međubrok (14:30):

- **\*\*Rižin krekeri s niskomasnim sirom\*\***
  - 20 g rižinih krekeri (76 kcal, 0.4 g masti)
  - 50 g niskomasnog sira (40 kcal, 1 g masti)

#### Večera (17:00):

- **\*\*Minestrone juha s puretinom\*\***

- 100 g purećih prsa na kockice (110 kcal, 1.2 g masti)

- 200 g povrća (mrkva, luk, celer, rajčica) u juhi (70 kcal, 0.5 g masti)

- 150 ml povrtnog temeljca (15 kcal, 0 g masti)

#### Međubrok (19:30):

- **\*\*Jogurt s niskim udjelom masti\*\***

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)

#### Ukupni dnevni unos masti: 16.8 grama

## **Prilogu 6. Inženjering upita s 8 pokušaja**

*Sistemska poruka:*

*Ti si vrhunski nutricionist sa više do 10 godina iskustva u kreiranju jelovnika za osobe s VLCAD poremećajem. Kreiraj dnevni jelovnik za osobu koja ima VLCAD poremećaj u prehrani u kojem dnevni unos masti nije veći od 20 grama. Napiši tok misli korak po korak prilikom kreiranja jelovnika.*

*Korisnik: Primjer dnevnog jelovnika na bazi puretine:*

##### Doručak (7:00):

- Griz s Kraš ekspresom

- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)

- 40 g griza (146 kcal, 1 g masti)

- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

##### Međubrok (9:30):

- Banana (120 g, 105 kcal, 0.3 g masti)

##### Ručak (12:00):

- Pureća prsa s povrćem



- 150 g purećih prsa (165 kcal, 1.8 g masti)

- 200 g kuhanog povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)

##### Međubrok (14:30):

- Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)

##### Večera (17:00):

- Omlet od bjelanjaka s povrćem

- 4 bjelanjka (68 kcal, 0.2 g masti)

- 150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)

##### Međubrok (19:30):

- Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)

Asistent: Možeš li mi dati još jedan primjer na bazi piletine?

Korisnik: Primjer dnevnog jelovnika na bazi piletine:

##### Doručak (7:00):

- Griz s Kraš ekspresom

- 200 ml mlijeka 1.5% mm (92 kcal, 3 g masti)

- 40 g griza (146 kcal, 1 g masti)

- 10 g Kraš eksprese (40 kcal, 0.3 g masti)

##### Međubrok (9:30):

- Banana (120 g, 105 kcal, 0.3 g masti)

##### Ručak (12:00):

- Piletina s rižom i povrćem

- 100 g pilećih prsa na žaru (120 kcal, 1.5 g masti)

- 100 g kuhane riže (130 kcal, 0.3 g masti)

- 150 g kuhanog povrća (mrkva, grašak) (70 kcal, 0.4 g masti)

##### *Međuobrok (14:30):*

- *Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)*

##### *Večera (17:00):*

- *Omlet od bjelanjaka s povrćem*

- *4 bjelanjaka (68 kcal, 0.2 g masti)*

- *150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)*

##### *Međuobrok (19:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

...

*Asistent: Možeš li mi dati još jedan primjer na bazi teletine?*

*Korisnik: Primjer dnevnog jelovnika na bazi teletine:*

##### *Doručak (7:00):*

- *Palenta s jogurtom*

- *50 g palente (167 kcal, 0.5 g masti)*

- *200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)*

##### *Međuobrok (9:30):*

- *Jabuka (150 g, 78 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Kuhana teletina s povrćem*

- *150 g telećeg buta (210 kcal, 8 g masti)*

- *200 g kuhanog povrća (špinat, brokula, mrkva) (85 kcal, 1 g masti)*

##### *Međuobrok (14:30):*

- *Rižin krekeri (20 g, 76 kcal, 0.4 g masti) s niskomasnim sirom (50 g, 40 kcal, 1 g masti)*

##### *Večera (17:00):*

- *Omlet od bjelanjaka s povrćem*

- 4 bjelanjaka (68 kcal, 0.2 g masti)

- 150 g povrća (paprika, tikvica) (60 kcal, 0.3 g masti)

##### *Međuobrok (19:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

*Asistent: Možeš li mi dati još jedan primjer?*

*Korisnik: Primjer dnevnog jelovnika na bazi ribe:*

##### *Doručak (7:00):*

- *Palenta s jogurtom*

- 50 g palente (167 kcal, 0.5 g masti)

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)

##### *Međuobrok (9:30):*

- *Jabuka (150 g, 78 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Pečena riba s povrćem*

- 100 g pečene bijele ribe (82 kcal, 1 g masti)

- 200 g kuhanog povrća (brokula, mrkva, cvjetača) (100 kcal, 1 g masti)

##### *Međuobrok (14:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

##### *Večera (17:00):*

- *Pečeni bakalar s povrćem*

- 100 g bakalara (82 kcal, 1 g masti)

- 200 g kuhanog povrća (brokula, mrkva, cvjetača) (100 kcal, 1 g masti)

##### *Međuobrok (19:30):*

- *Mandarina (140 g, 65 kcal, 0.2 g masti)*

*Asistent: Možeš li mi dati još jedan primjer?*

*Korisnik: Primjer dnevnog jelovnika:*

##### *Doručak (7:00):*

- *Palenta s jogurtom*

- *50 g palente (167 kcal, 0.5 g masti)*

- *200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)*

##### *Međuobrok (9:30):*

- *Jabuka (150 g, 78 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Pečena piletina s pireom od krumpira*

- *100 g pečene piletkine (165 kcal, 3.6 g masti)*

- *150 g pire krumpira s maslinovim uljem (120 kcal, 3 g masti)*

##### *Međuobrok (14:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

##### *Večera (17:00):*

- *Pečena riba s povrćem*

- *100 g pečene bijele ribe (82 kcal, 1 g masti)*

- *200 g kuhanog povrća (brokula, mrkva, cvjetača) (100 kcal, 1 g masti)*

##### *Međuobrok (19:30):*

- *Mandarina (140 g, 65 kcal, 0.2 g masti)*

*Asistent: Možeš li mi dati još jedan primjer?*

*Korisnik: Evo još jednog primjera dnevnog jelovnika:*

##### *Doručak (7:00):*

- *Palenta s jogurtom*

- *50 g palente (167 kcal, 0.5 g masti)*

- *200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)*

##### *Međuobrok (9:30):*

- *Jabuka (150 g, 78 kcal, 0.3 g masti)*

##### *Ručak (12:00):*

- *Bolonjez s tjesteninom*

- *100 g mljevene puretine (130 kcal, 3.3 g masti)*

- *150 g kuhane tjestenine (200 kcal, 1 g masti)*

- *100 g pelata (25 kcal, 0.2 g masti)*

##### *Međuobrok (14:30):*

- *Jogurt s niskim udjelom masti (200 ml, 100 kcal, 3 g masti)*

##### *Večera (17:00):*

- *Pečeni bakalar s povrćem*

- *100 g bakalara (82 kcal, 1 g masti)*

- *200 g kuhanog povrća (brokula, mrkva, cvjetača) (100 kcal, 1 g masti)*

##### *Međuobrok (19:30):*

- *Mandarina (140 g, 65 kcal, 0.2 g masti)*

*Asistent: Možeš li mi dati još jedan primjer?*

*Korisnik: Evo još jedan primjer dnevnog jelovnika:*

**\*\*Doručak (7:30)\*\*:** *Griz s Kraš Expressom*

- **\*\*Griz\*\*** (50g): *175 kcal, 1.1g masti*

- **\*\*Kraš Express** (2 žlice)\*\*: *70 kcal, 1.2g masti*

- **\*\*Mliječna krava 1.5% masti** (200ml)\*\*: *90 kcal, 3g masti*

- **Ukupno**: 335 kcal, 5.3g masti

**Međubrok 1 (10:00)**: Jogurt s niskim udjelom masti

- **Jogurt (150g, 0.9% masti)**: 53 kcal, 1.4g masti

**Ručak (12:30)**: Pečeni oslić sa krumpirom i kuhanim povrćem

- **Oslić (100g)**: 85 kcal, 0.7g masti

- **Krumpir** (150g): 150 kcal, 0.3g masti

- **Kuhano povrće (mješavina, 150g)**: 70 kcal, 0.5g masti

- **Ukupno**: 305 kcal, 1.5g masti

**Međubrok 2 (15:00)**: Rižini kolačići (3 komada)

- **Rižin kolač (10g)**: 35 kcal, 0.2g masti (3 komada: 105 kcal, 0.6g masti)

**Večera (18:00)**: Pileća salata (pileća prsa, kuhano povrće)

- **Pileća prsa** (100g): 165 kcal, 3.6g masti

- **Kuhano povrće (150g)**: 70 kcal, 0.5g masti

- **Ukupno**: 235 kcal, 4.1g masti

**Međubrok 3 (20:30)**: Krekeri s nemasnim sirom

- **Krekeri (30g)**: 120 kcal, 1g masti

- **Nemasni sir (50g)**: 55 kcal, 1g masti

- **Ukupno**: 175 kcal, 2g masti

**Ukupno za dan**: 1208 kcal, 14.9g masti

Asistent: Možeš li mi dati još jedan primjer?

Korisnik: Evo još jedan primjer dnevnog jelovnika:

**Doručak (7:30)**: Palenta s jogurtom

Po uzoru na ponedjeljak:

- **Ukupno**: 125 kcal, 4.8g masti

**\*\*Međuobrok 1 (10:00)\*\*:** *Naranča*

- **\*\*Naranča\*\*** (130g): 62 kcal, 0.2g masti

**\*\*Ručak (12:30)\*\*:** *Pileća prsa s pirjanom rižom i kuhanim povrćem*

- **\*\*Pileća prsa\*\*** (100g): 165 kcal, 3.6g masti

- **\*\*Pirjana riža\*\*** (100g): 130 kcal, 0.5g masti

- **\*\*Kuhano povrće (150g)\*\***: 70 kcal, 0.5g masti

- **\*\*Ukupno\*\***: 365 kcal, 4.6g masti

**\*\*Međuobrok 2 (15:00)\*\*:** *Jogurt s voćem*

- **\*\*Jogurt (150g, 0.9% masti)\*\***: 53 kcal, 1.4g masti

- **\*\*Voće (banane, jagode, 100g)\*\***: 50 kcal, 0.2g masti

- **\*\*Ukupno\*\***: 103 kcal, 1.6g masti

**\*\*Večera (18:00)\*\*:** *Špageti s umakom od rajčice*

- **\*\*Špageti\*\*** (100g): 150 kcal, 0.9g masti

- **\*\*Umak od rajčice\*\*** (100g): 50 kcal, 0.5g masti

- **\*\*Ukupno\*\***: 200 kcal, 1.4g masti

**\*\*Međuobrok 3 (20:30)\*\*:** *Svježi sir s medom*

- **\*\*Svježi sir (50g)\*\***: 55 kcal, 1g masti

- **\*\*Med (1 žlica)\*\***: 64 kcal, 0g masti

- **\*\*Ukupno\*\***: 119 kcal, 1g masti

**\*\*Ukupno za dan\*\***: 974 kcal, 14.6g masti

*Korisnik: Kreiraj mi dnevni jelovnik.*

## **Prilog 7. Dnevni jelovnik: Inženjering upita s 8 pokušaja**

*Naravno, evo jednog primjera dnevnog jelovnika koji je prilagođen osobama s VLCAD poremećajem, s dnevnim unosom masti ne većim od 20 grama:*

**### Doručak (7:30)**

***\*\*Zobena kaša s voćem\*\****

- 40 g zobenih pahuljica (150 kcal, 3 g masti)
- 200 ml mlijeka s niskim udjelom masti (92 kcal, 3 g masti)
- 100 g svježih borovnica (57 kcal, 0 g masti)
- ***\*\*Ukupno\*\****: 299 kcal, 6 g masti

***### Međubrok (10:00)***

***\*\*Voćna salata\*\****

- 150 g jabuka (78 kcal, 0.3 g masti)
- 100 g naranče (47 kcal, 0.2 g masti)
- ***\*\*Ukupno\*\****: 125 kcal, 0.5 g masti

***### Ručak (12:30)***

***\*\*Pileći file s kvinojom i povrćem\*\****

- 100 g pilećih prsa na žaru (165 kcal, 3.6 g masti)
- 150 g kuhane kvinoje (120 kcal, 1.9 g masti)
- 200 g povrća (brokula, mrkva, cvjetača, kuhana) (85 kcal, 1 g masti)
- ***\*\*Ukupno\*\****: 370 kcal, 6.5 g masti

***### Međubrok (15:00)***

***\*\*Prirodni nezaslađeni jogurt\*\****

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)
- ***\*\*Ukupno\*\****: 100 kcal, 3 g masti

***### Večera (18:00)***

***\*\*Pečeni bakalar s povrćem\*\****

- 150 g bakalara pečenog (123 kcal, 1.5 g masti)
- 200 g kuhanog povrća (špinat, brokula, šparoge) (70 kcal, 0.5 g masti)



- **Ukupno**: 193 kcal, 2 g masti

### Međubrok (20:30)

**Rižin krekeri s niskomasnim sirom**

- 20 g rižinih krekeri (76 kcal, 0.4 g masti)

- 50 g niskomasnog sira (40 kcal, 1 g masti)

- **Ukupno**: 116 kcal, 1.4 g masti

### Ukupno za dan:

**Ukupne kalorije**: 1203 kcal

**Ukupne masnoće**: 19.4 g masti

Ovo bi trebao biti dobro izbalansiran plan prehrane koji zadovoljava sve kalorijske i nutritivne potrebe osobe s VLCAD poremećajem i drži dnevni unos masnoće ispod 20 grama.

## **Prilog 8. Upit: Tok misli korak po korak**

Prilikom kreiranja jelovnika, potrebno je zadovoljiti sljedeće korake:

*Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika*

1. Unos masti: Ograničenje unosa masti na ne više od 20 grama dnevno.
2. Unos ugljikohidrata: Osoba treba dovoljno ugljikohidrata kako bi imala dovoljnu količinu energije.
3. Unos proteina: Proteini su također važni za održavanje mišićne mase i funkcije tijela.
4. Unos mikronutrijenata: Osigurati dovoljan unos vitamina i minerala kroz raznoliku prehranu.

*Korak 2: Odabir pravilnih namirnica*

1. Nemasne proteinske namirnice: Pileća prsa, pureća prsa, teleći but, riba (posebno bijela riba), bjelanjci, nemasni svježi sir, leća, grah, grašak, kvinoja.
2. Ugljikohidrati: Zob, kvinoja, heljda, proso, krumpir, riža, tjestenina, banane.
3. Zamjene za masti: MCT ulje (ulje s masnim kiselinama srednjeg lanca) kao dodatak prehrani, s obzirom na to da se lakše metabolizira kod osoba s VLCAD.

*Korak 3: Praćenje unosa masti:*

- *Točno evidentirati i kontrolirati unos masti.*

#### *Korak 4: Izrada jelovnika*

##### *Primjer jednog jelovnika:*

###### *### Doručak (7:30)*

###### *\*\*Zobena kaša s voćem\*\**

- *40 g zobenih pahuljica (150 kcal, 3 g masti)*
- *200 ml mlijeka s niskim udjelom masti (92 kcal, 3 g masti)*
- *100 g svježih borovnica (57 kcal, 0 g masti)*
- *\*\*Ukupno\*\*:* 299 kcal, 6 g masti

###### *### Međubrok (10:00)*

###### *\*\*Voćna salata\*\**

- *150 g jabuka (78 kcal, 0.3 g masti)*
- *100 g naranče (47 kcal, 0.2 g masti)*
- *\*\*Ukupno\*\*:* 125 kcal, 0.5 g masti

###### *### Ručak (12:30)*

###### *\*\*Pileći file s kvinojom i povrćem\*\**

- *100 g pilećih prsa na žaru (165 kcal, 3.6 g masti)*
- *150 g kuhane kvinoje (120 kcal, 1.9 g masti)*
- *200 g povrća (brokula, mrkva, cvjetača, kuhana) (85 kcal, 1 g masti)*
- *\*\*Ukupno\*\*:* 370 kcal, 6.5 g masti

###### *### Međubrok (15:00)*

###### *\*\*Prirodni nezaslađeni jogurt\*\**

- *200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)*
- *\*\*Ukupno\*\*:* 100 kcal, 3 g masti

#### Večera (18:00)

***\*\*Pečeni bakalar s povrćem\*\****

- 150 g bakalara pečenog (123 kcal, 1.5 g masti)

- 200 g kuhanog povrća (špinat, brokula, šparoge) (70 kcal, 0.5 g masti)

- ***\*\*Ukupno\*\****: 193 kcal, 2 g masti

#### Međubrok (20:30)

***\*\*Rižin krekeri s niskomasnim sirom\*\****

- 20 g rižinih krekeri (76 kcal, 0.4 g masti)

- 50 g niskomasnog sira (40 kcal, 1 g masti)

- ***\*\*Ukupno\*\****: 116 kcal, 1.4 g masti

#### Ukupno za dan:

***\*\*Ukupne kalorije\*\****: 1203 kcal

***\*\*Ukupne masnoće\*\****: 19.4 g masti

## **Prilog 9. Dnevni jelovnik: Tok misli korak po korak**

#### Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika

1. ***\*\*Unos masti\*\****: Ograničiti unos masti na ne više od 20 grama dnevno.
2. ***\*\*Unos ugljikohidrata\*\****: Osigurati dovoljnu količinu ugljikohidrata za energiju.
3. ***\*\*Unos proteina\*\****: Osigurati dovoljan unos proteina za održavanje mišićne mase i funkcije tijela.
4. ***\*\*Unos mikronutrijenata\*\****: Osigurati dovoljan unos vitamina i minerala kroz raznoliku prehranu.

#### Korak 2: Odabir pravilnih namirnica

1. ***\*\*Nemasne proteinske namirnice\*\****:

- Teletina (osobito nemasni dijelovi kao file ili but)

- Drugi nemasni izvori proteina poput bjelanjaka i nemasnog sira.

2. **Ugljikohidrati**:

- Zob, kvinoja, heljda, proso, krumpir, riža, tjestenina, banane.

3. **Zamjene za masti**:

- MCT ulje kao dodatak prehrani.

### Korak 3: Praćenje unosa masti

- Točno evidentirati i kontrolirati unos masti.

### Korak 4: Izrada jelovnika na bazi teletine

### Doručak (7:30)

**Omljet od bjelanjaka s povrćem**

- 4 bjelanjka (68 kcal, 0 g masti)

- 100 g špinata (23 kcal, 0.3 g masti)

- 100 g rajčica (18 kcal, 0.2 g masti)

- 1 žlica MCT ulja (115 kcal, 13 g masti)

- **Ukupno**: 224 kcal, 13.5 g masti

### Međubrok (10:00)

**Jabuka i mrkva**

- 150 g jabuka (78 kcal, 0.3 g masti)

- 100 g mrkva (41 kcal, 0.2 g masti)

- **Ukupno**: 119 kcal, 0.5 g masti

### Ručak (12:30)

**Teletina na žaru s kvinojom i povrćem**

- 150 g telećih prsa pečениh (225 kcal, 4 g masti)

- 150 g kuhane kvinoje (120 kcal, 1.9 g masti)

- 200 g povrća (brokula, mrkva, cvjetača, kuhana) (85 kcal, 1 g masti)

- **\*\*Ukupno\*\***: 430 kcal, 6.9 g masti

### Međubrok (15:00)

**\*\*Prirodni nezaslađeni jogurt\*\***

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)

- **\*\*Ukupno\*\***: 100 kcal, 3 g masti

### Večera (18:00)

**\*\*Pečena teletina s povrćem\*\***

- 150 g pečene teletine (225 kcal, 4 g masti)

- 200 g kuhanog povrća (špinat, brokula, šparoge) (70 kcal, 0.5 g masti)

- **\*\*Ukupno\*\***: 295 kcal, 4.5 g masti

### Međubrok (20:30)

**\*\*Rižin krekeri s niskomasnim sirom\*\***

- 20 g rižinih krekeri (76 kcal, 0.4 g masti)

- 50 g niskomasnog sira (40 kcal, 1 g masti)

- **\*\*Ukupno\*\***: 116 kcal, 1.4 g masti

### Ukupno za dan

**\*\*Ukupne kalorije\*\***: 1284 kcal

**\*\*Ukupne masnoće\*\***: 19.8 g masti

**\*\*Napomene\*\***: Omlet za doručak s četiri bjelanjka pruža dobru količinu proteina bez dodatka masti, dok MCT ulje dodaje potrebne masti koje se lakše metaboliziraju za osobe s VLCAD poremećajem. Teletina služi kao glavni izvor proteina za ručak i večeru, uz nemasne priloge poput kvinoje i raznog povrća. Međubroci su lagani, ali nutritivno bogati, što omogućava da se ukupni unos masnoće zadrži unutar preporučenih granica.

## **Prilog 10. Poboljšani upit: Tok misli korak po korak**

Prilikom kreiranja jelovnika, potrebno je zadovoljiti sljedeće korake:

*Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika*

- 1. Unos masti: Ograničenje unosa masti na ne više od 20 grama dnevno.*
- 2. Unos ugljikohidrata: Osoba treba dovoljno ugljikohidrata kako bi imala dovoljnu količinu energije.*
- 3. Unos proteina: Proteini su također važni za održavanje mišićne mase i funkcije tijela.*
- 4. Unos mikronutrijenata: Osigurati dovoljan unos vitamina i minerala kroz raznoliku prehranu.*

*Korak 2: Odabir pravilnih namirnica*

- 1. Nemasne proteinske namirnice: Pileća prsa, pureća prsa, teleći but, riba (posebno bijela riba), bjelanjci, nemasni svježi sir, leća, grah, grašak, kvinoja.*
- 2. Ugljikohidrati: Zob, kvinoja, heljda, proso, krumpir, riža, tjestenina, banane.*
- 3. Zamjene za masti: MCT ulje (ulje s masnim kiselinama srednjeg lanca) kao dodatak prehrani, s obzirom na to da se lakše metabolizira kod osoba s VLCAD.*

*Korak 3: Vrijeme pripreme obroka:*

- Uzmi u obzir da osoba ide u osnovnu školu, stoga obroci trebaju biti jednostavni i brzi za pripremiti (maksimalno 60 minuta)*

*Korak 4: Praćenje unosa masti:*

- Točno evidentirati i kontrolirati unos masti.*

*Korak 5: Izrada jelovnika*

*Primjer jednog jelovnika:*

*### Doručak (7:30)*

*\*\*Zobena kaša s voćem\*\**

- 40 g zobenih pahuljica (150 kcal, 3 g masti)*
- 200 ml mlijeka s niskim udjelom masti (92 kcal, 3 g masti)*
- 100 g svježih borovnica (57 kcal, 0 g masti)*
- \*\*Ukupno\*\*: 299 kcal, 6 g masti*

*### Međubrok (10:00)*

***\*\*Voćna salata\*\****

- 150 g jabuka (78 kcal, 0.3 g masti)
- 100 g naranče (47 kcal, 0.2 g masti)
- ***\*\*Ukupno\*\****: 125 kcal, 0.5 g masti

***### Ručak (12:30)***

***\*\*Pileći file s kvinojom i povrćem\*\****

- 100 g pilećih prsa na žaru (165 kcal, 3.6 g masti)
- 150 g kuhane kvinoje (120 kcal, 1.9 g masti)
- 200 g povrća (brokula, mrkva, cvjetača, kuhana) (85 kcal, 1 g masti)
- ***\*\*Ukupno\*\****: 370 kcal, 6.5 g masti

***### Međubrok (15:00)***

***\*\*Prirodni nezaslađeni jogurt\*\****

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)
- ***\*\*Ukupno\*\****: 100 kcal, 3 g masti

***### Večera (18:00)***

***\*\*Pečeni bakalar s povrćem\*\****

- 150 g bakalara pečenog (123 kcal, 1.5 g masti)
- 200 g kuhanog povrća (špinat, brokula, šparoge) (70 kcal, 0.5 g masti)
- ***\*\*Ukupno\*\****: 193 kcal, 2 g masti

***### Međubrok (20:30)***

***\*\*Rižin krekeri s niskomasnim sirom\*\****

- 20 g rižinih krekeri (76 kcal, 0.4 g masti)
- 50 g niskomasnog sira (40 kcal, 1 g masti)
- ***\*\*Ukupno\*\****: 116 kcal, 1.4 g masti

#### Ukupno za dan:

**\*\*Ukupne kalorije\*\***: 1203 kcal

**\*\*Ukupne masnoće\*\***: 19.4 g masti

### **Prilogu 11. Dnevni jelovnik: Poboljšani upit s tokom misli korak po korak**

## Tok misli pri kreiranju jelovnika za osobu s VLCAD poremećajem

#### Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika

1. **\*\*Unos masti\*\***:

- Ograničiti unos masti na ne više od 20 grama dnevno.

2. **\*\*Unos ugljikohidrata\*\***:

- Osigurati dovoljnu količinu ugljikohidrata za potrebnu energiju.

3. **\*\*Unos proteina\*\***:

- Osigurati dovoljan unos proteina za održavanje mišićne mase i funkcije tijela.

4. **\*\*Unos mikronutrijenata\*\***:

- Osigurati dovoljan unos vitamina i minerala kroz raznoliku prehranu.

#### Korak 2: Odabir pravilnih namirnica

1. **\*\*Nemasne proteinske namirnice\*\***:

- Teleća prsa, pileća prsa, pureća prsa, riba (posebno bijela riba), bjelanjci, nemasni sir, leća, grah, grašak, kvinoja.

2. **\*\*Ugljikohidrati\*\***:

- Zob, kvinoja, heljda, proso, krumpir, riža, tjestenina, banane.

3. **\*\*Zamjene za masti\*\***:

- MCT ulje (ulje s masnim kiselinama srednjeg lanca) zbog lakše metabolizacije kod osoba s VLCAD.

#### Korak 3: Vrijeme pripreme obroka:



- Uzeti u obzir da osoba ide u osnovnu školu, stoga obroci trebaju biti jednostavni i brzi za pripremu (maksimalno 60 minuta).

### Korak 4: Praćenje unosa masti:

- Točno evidentirati i kontrolirati unos masti.

### Korak 5: Izrada jelovnika

### Doručak (7:00)

**\*\*Zobena kaša s voćem i medom\*\***

- 40 g zobenih pahuljica (150 kcal, 3 g masti)

- 200 ml mlijeka s niskim udjelom masti (92 kcal, 3 g masti)

- 100 g borovnica (57 kcal, 0 g masti)

- 1 čajna žličica meda (21 kcal, 0 g masti)

- **\*\*Ukupno\*\***: 320 kcal, 6 g masti

### Međubrok (10:00)

**\*\*Voćna salata\*\***

- 150 g jabuka (78 kcal, 0.3 g masti)

- 100 g naranče (47 kcal, 0.2 g masti)

- **\*\*Ukupno\*\***: 125 kcal, 0.5 g masti

### Ručak (12:30)

**\*\*Teleća prsa s kvinojom i povrćem\*\***

- 100 g telećih prsa na žaru (150 kcal, 4 g masti)

- 150 g kuhane kvinoje (120 kcal, 1.9 g masti)

- 200 g povrća (brokula, mrkva, cvjetača, kuhana) (85 kcal, 1 g masti)

- **\*\*Ukupno\*\***: 355 kcal, 6.9 g masti

### Međubrok (15:00)

**\*\*Prirodni nezaslađeni jogurt\*\***

- 200 ml jogurta s niskim udjelom masti (100 kcal, 3 g masti)

- **Ukupno**: 100 kcal, 3 g masti

### Večera (18:00)

**Pečeni bakalar s povrćem**

- 150 g bakalara pečenog (123 kcal, 1.5 g masti)

- 200 g šparoga (42 kcal, 0.5 g masti)

- 100 g pečenog krumpira bez masnoće (93 kcal, 0.1 g masti)

- **Ukupno**: 258 kcal, 2.1 g masti

### Međubrok (20:30)

**Rižin krekeri s niskomasnim sirom**

- 20 g rižinih krekeri (76 kcal, 0.4 g masti)

- 50 g niskomasnog sira (40 kcal, 1 g masti)

- **Ukupno**: 116 kcal, 1.4 g masti

### Ukupno za dan:

- **Ukupne kalorije**: 1274 kcal

- **Ukupne masnoće**: 19.9 g masti

### Razrada obroka:

1. **Doručak**:

- Priprema zobene kaše je jednostavna i traje manje od 10 minuta. Odabir voća i meda poboljšava okus i dodaje mikronutrijente.

2. **Međubrok 1**:

- Voćna salata je praktična za školu. Lako se priprema večer prije i bogata je vitaminima.

3. **Ručak**:

- Teleća prsa na žaru (mogu se pripremiti prethodnu večer i lako zagrijati), kvinoju kuhati 15-20 minuta, povrće može biti brzo kuhano na pari.

4. **\*\*Međuobrok 2\*\***:

- Jogurt je jednostavan i brz izbor, osigurava proteine i kalcij.

5. **\*\*Večera\*\***:

- Bakalar se peče oko 15-20 minuta, povrće može biti kuhano na pari, a krumpir pečen u pećnici bez dodatnih masti.

6. **\*\*Međuobrok 3\*\***:

- Rižini krekeri s niskomasnim sirom su brzi za pripremiti i praktični za večernji obrok.

### Zaključak:

- Ovaj jelovnik je koncipiran tako da zadovoljava sve nutritivne potrebe korisnika sa VLCAD poremećajem, ograničava unos masti, a istovremeno pruža sve potrebne makro- i mikronutrijente. Svi obroci su jednostavni i brzi za pripremu, uzimajući u obzir školsku obavezu djeteta.

## **Prilogu 12. ReART upit**

*Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika*

*Pojašnjenje: Jasno formulirajte specifične nutritivne potrebe osobe radi boljeg razumijevanja. Koje su osnovne nutritivne potrebe (unos masti, ugljikohidrati, proteini, vitamini i minerali)?*

*Zaključak upita: Ograničiti unos masti na ne više od 20 grama dnevno, osigurati adekvatan unos proteina i ugljikohidrata, te osigurati raznolik unos mikronutrijenata (vitamina i minerala).*

*Korak 2: Odabir pravilnih namirnica*

*Pojašnjenje: Navesti koje su namirnice prikladne za unos proteina, ugljikohidrata te kontrolu masnoće.*

- *Koje nemasne proteinske namirnice možemo koristiti?*
- *Koji izvori ugljikohidrata su najbolji za osiguranje energije?*
- *Koje zamjene za masti možemo uključiti u jelovnik?*

*Zaključak upita: Odabrane namirnice trebaju biti bogate proteinima i ugljikohidratima, a istovremeno imati nizak udio masti. Dodatak MCT ulja kao zamjene za tradicionalne masti može pomoći u lakšoj razgradnji masti kod osobe s VLCAD poremećajem.*

### *Korak 3: Vrijeme pripreme obroka*

*Pojašnjenje: Pojasnite zahtjeve za vremensku učinkovitost pripreme obroka.*

- Koliko vremena bi priprema svakog obroka trebalo trajati?*

*Zaključak upita: Obroci trebaju biti jednostavni i brzi za pripremu, maksimalno 60 minuta.*

### *Korak 4: Praćenje unosa masti*

*Pojašnjenje: Kako točno pratiti i evidentirati unos masti?*

- Koji alati ili metode se mogu koristiti za točno praćenje unosa masti?*

*Zaključak upita: Evidentirati i kontrolirati svaki unos masti unutar obroka.*

### *Korak 5: Izrada jelovnika*

*Pojašnjenje: Kreiraj primjer jelovnika na temelju gore navedenih koraka.*

- Kako strukturirati obroke tijekom dana kako bi zadovoljili sve nutritivne kriterije i vremenske zahtjeve?*

## **Prilogu 13. Dnevni jelovnik ReART**

*Da bi se izradio optimalan dnevni jelovnik po gore navedenim koracima, važno je pratiti redoslijed kako bismo osigurali sve nutritivne potrebe, odabrali prikladne namirnice, osigurali učinkovitu pripremu i kontrolirali unos masti. Evo kako bi to moglo izgledati:*

*### Korak 1: Identifikacija nutritivnih potreba korisnika jelovnika*

*Nutritivne potrebe uključuju:*

- Maksimalno 20 grama masti dnevno*
- Adekvatni unos proteina i ugljikohidrata*
- Raznolik unos vitamina i minerala*

*### Korak 2: Odabir pravilnih namirnica*

*##### Nemasne proteinske namirnice:*

- Piletina (bez kože)*
- Puretina (bez kože)*

- *Bijela riba (bakalar, oslić)*

- *Tofu*

- *Jaja (posebno bjelanjci)*

##### *Izvori ugljikohidrata:*

- *Smeđa riža*

- *Kvinoja*

- *Integralni kruh*

- *Zob*

- *Voće (jabuke, banane, bobičasto voće)*

- *Povrće (brokula, mrkva, špinat)*

##### *Zamjene za masti:*

- *MCT ulje (srednjelančani trigliceridi)*

- *Koristiti niskomasne ili nemasne verzije hrane kad god je moguće*

#### *Korak 3: Vrijeme pripreme obroka*

*S obzirom na to da obroci trebaju biti gotovi unutar 60 minuta, usredotočit ćemo se na jednostavne recepte i tehnike pripreme poput kuhanja, pečenja u pećnici i prženja na minimalnoj količini MCT ulja.*

#### *Korak 4: Praćenje unosa masti*

*Kako biste točno pratili unos masti, predlaže se korištenje aplikacija poput MyFitnessPal ili Cronometer koje omogućuju detaljno praćenje nutritivnih vrijednosti unesene hrane.*

#### *Korak 5: Izrada jelovnika*

##### *Doručak*

*\*\*Zobena kaša s bobičastim voćem\*\**

- *50 g zobenih pahuljica (kuhanih u vodi)*

- *100 g bobičastog voća (maline, borovnice)*

- 1 žličica MCT ulja (pridodano poslije kuhanja)

**\*\*Nutritivne vrijednosti:\*\***

- Proteini: 5-10 g

- Ugljikohidrati: 35-40 g

- Masti: 5 g

##### Međubrok

**\*\*Grčki jogurt s jabukom\*\***

- 200 g nemasnog grčkog jogurta

- 1 srednja jabuka, narezana

**\*\*Nutritivne vrijednosti:\*\***

- Proteini: 10-15 g

- Ugljikohidrati: 20-25 g

- Masti: 0-2 g

##### Ručak

**\*\*Pečena piletina s brokulom i kvinojom\*\***

- 150 g pilećih prsa (pečena bez kože, uz minimalno MCT ulja)

- 100 g brokule (kuhane na pari)

- 50 g kvinoje (kuhane)

**\*\*Nutritivne vrijednosti:\*\***

- Proteini: 30-35 g

- Ugljikohidrati: 40-45 g

- Masti: 5-7 g

##### Međubrok

**\*\*Šaka badema i banana\*\***

- 15 g badema

- 1 mala banana

**\*\*Nutritivne vrijednosti:\*\***

- Proteini: 3-5 g

- Ugljikohidrati: 30-35 g

- Masti: 10 g

##### Večera

**\*\*Bakalar s pečenim povrćem\*\***

- 150 g pečenog bakalara (uz minimalno MCT ulja)

- 200 g pečenog povrća (mrkva, tikvice, paprike)

**\*\*Nutritivne vrijednosti:\*\***

- Proteini: 30-35 g

- Ugljikohidrati: 20-25 g

- Masti: 3-5 g

##### Ukupne dnevne nutritivne vrijednosti:

- Proteini: 78-105 g

- Ugljikohidrati: 145-170 g

- Masti: 20 g

#### Zaključak

*Ovaj jelovnik ispunjava sve nutritivne ciljeve za dan:*

- Maksimalno 20 g masti

- Adekvatni unos proteina i ugljikohidrata

- Raznovrsne namirnice osiguravaju bogat unos vitamina i minerala

- Vrijeme pripreme svih obroka unutar zadatka (maksimalno 60 minuta)

*S ovim koracima i primjerom jelovnika, korisnik može lakše pratiti svoje nutritivne potrebe i osigurati balansirane obroke uz kontrolirani unos masti.*

**Priloženi disk sadržava**

1. Diplomski rad „Inženjering upita za velike jezične modele“ u .docx formatu
2. Diplomski rad „Inženjering upita za velike jezične modele“ u .pdf formatu

---

Potpis autora