УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Вук Стефановић

ЕКСПЕРИМЕНТАЛНА АНАЛИЗА УТИЦАЈА КОМПОНЕНТИ RAG СИСТЕМА НА КВАЛИТЕТ ОДГОВОРА ВЕЛИКИХ ЈЕЗИЧКИХ МОДЕЛА

мастер рад

Београд, 2026.

Ментор:

др Младен Николић, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Јелена Граовац, ванредни професор
Универзитет у Београду, Математички факултет

др Мирјана Маљковић Ружичић, доцент
Универзитет у Београду, Математички факултет

Датум одбране: _____

Наслов мастер рада: Експериментална анализа утицаја компоненти RAG система на квалитет одговора великих језичких модела

Резиме: Системи засновани на приступу генерисања појачаног претрагом (енг. Retrieval-Augmented Generation, RAG) комбинују претрагу докумената са великим језичким моделима како би генерисали одговоре поткрепљене конкретним изворима. Њихов квалитет, међутим, не одређује само генеративни модел, већ и низ компоненти које претходе генерацији одговора. Овај рад испитује колико избор тих компоненти — стратегије сегментације, модела за векторску репрезентацију текста, броја враћених сегмената и начина претраге — утиче на квалитет претраге и на квалитет коначних одговора.

Над корпусом од 326 историјских чланака са Википедије поређено је осам конфигурација RAG система, добијених поступним варирањем појединачних компоненти. Квалитет претраге мерен је метрикама Одзив@k и MAP, а квалитет одговора оценама тачности, утемељености, релевантности и потпуности, добијеним применом приступа LLM као судија и провереним мануелном евалуацијом на подскупу питања.

Налази показују да пажљив избор компоненти претраге значајно подиже квалитет система, при чему најбоље резултате остварује хибридна конфигурација са ренгирањем. Ипак, виши квалитет претраге не прати увек сразмерно бољи одговор, што указује на то да се квалитет RAG система не може поуздано проценити само на основу метрика претраге, већ захтева и непосредну анализу генерисаних одговора.

Кључне речи: велики језички модели, генерисање појачано претрагом, RAG, претрага информација, векторске репрезентације текста, евалуација одговора

Садржај

Глава 1

Увод

Последњих година велики језички модели постали су један од најзначајнијих праваца развоја вештачке интелигенције. Ови модели показују високу способност разумевања и генерисања природног језика, али њихова примена у задацима који захтевају тачне, проверљиве и документима поткрепљене одговоре и даље носи одређене изазове. Један од најважнијих проблема је појава халуцинација, односно генерисање тврдњи које делују уверљиво, али се не могу потврдити на основу релевантних извора.

Као начин за ублажавање овог проблема развијен је приступ генерисања појачаног претрагом (RAG). Уместо да се одговор формира искључиво на основу параметарског знања модела, RAG систем најпре проналази релевантне сегменте докумената, а затим их користи као додатни контекст приликом генерисања одговора. На тај начин се процес генерисања усмерава ка информацијама које су експлицитно присутне у релевантним документима, што може побољшати тачност и проверљивост одговора.

Ипак, квалитет RAG система не зависи само од генеративног модела. Значајан утицај имају и компоненте које претходе самој генерацији одговора: начин сегментације докумената, избор модела за векторску репрезентацију текста, број враћених сегмената и начин претраге релевантног контекста. Ако систем не пронађе довољно релевантан контекст, чак и квалитетан генеративни модел може дати непотпун или нетачан одговор.

Циљ овог рада је да се експериментално испита у којој мери промене ових компоненти утичу на квалитет проналажења релевантног контекста и на квалитет коначних одговора.

Експерименти су спроведени над корпусом од 326 историјских чланака

прикупљених са Википедије. Поређено је осам различитих конфигурација RAG система. Неке од конфигурација користе претрагу на основу векторских репрезентација (енг. dense retrieval), која релевантне сегменте проналази на основу векторских репрезентација текста, док друге користе хибридни приступ, који претрагу на основу векторских репрезентација комбинује са лексичком, уз фузију и ренгирање резултата.

Квалитет претраге процењен је помоћу метрика Одзив@k и MAP, док је квалитет генерисаних одговора анализиран на основу четири критеријума: тачност, утемељеност у контексту, релевантност и потпуност. Поред евалуације применом приступа LLM као судија, део резултата је додатно проверен мануелном евалуацијом.

Резултати показују да појединачне одлуке у дизајну RAG система могу значајно утицати на квалитет коначних одговора. У појединим случајевима уочено је да боље вредности метрика претраге не доводе нужно до сразмерног побољшања квалитета генерисаних одговора. Због тога се у раду посебно анализира однос између квалитета претраге и квалитета генерисаних одговора, као и типичне грешке које се јављају код различитих типова питања.

Глава 2

Теоријске основе

У овом поглављу изложени су теоријски концепти неопходни за разумевање система за генерисање одговора уз претрагу докумената. Најпре се разматрају велики језички модели, њихове основне карактеристике и ограничења. Након тога уводи се приступ генерисања појачаног претрагом (RAG) као један од начина за повезивање генеративних модела са спољашњим изворима знања. У наставку се обрађују кључне компоненте RAG система: сегментација докумената, векторске репрезентације текста, методе претраге и поступци евалуације.

2.1 Велики језички модели

Велики језички модели (енг. Large Language Models, LLM) представљају класу модела дубоког учења обучених на великим количинама текстуалних података са циљем моделовања законитости природног језика. У контексту савремених генеративних система посебно су значајни ауторегресивни језички модели, код којих се текст генерише постепеним предвиђањем наредног токена на основу претходног контекста. Један од познатих примера таквог модела је GPT-3 [1]. Токен може представљати реч, део речи или знак интерпункције.

Токенизација

Језички модели не раде директно са текстом у његовом изворном облику. Пре уласка у модел, текст се дели на мање јединице, односно токене, а сваки

токен се затим повезује са одговарајућим идентификатором из речника модела. Овај поступак назива се токенизација (енг. tokenization).

Савремени језички модели најчешће користе токенизацију на нивоу делова речи (енг. subword tokenization). Овај приступ представља компромис између токенизације на нивоу карактера, која доводи до дугих секвенци, и токенизације на нивоу целих речи, која захтева велики речник и теже обухвата ретке или непознате речи. Честе речи могу бити представљене једним токеном, док се ређе или сложеније речи разлажу на мање делове. На пример, реч "unprecedented" може, у зависности од токенизатора, бити разложена на више токена, као што су "un", "pre" и "cedented". Овакав приступ омогућава моделу да ради и са речима које нису директно виђене током обучавања, уз контролисану величину речника.

Начин токенизације утиче на више аспеката језичког модела и система који га користи. Величина речника одређује колико различитих токена модел може да разликује. Већи речник омогућава да се већи број честих речи представи једним токеном, али истовремено повећава меморијске захтеве модела. Поред тога, исти текст може имати различиту дужину у токенима у зависности од примењеног токенизатора, што је посебно важно због ограничења контекстног прозора модела и поступка сегментације докумената у RAG системима.

Трансформер архитектура

Архитектурална основа савремених великих језичких модела је трансформер [15]. Трансформер уводи механизам пажње (енг. attention mechanism), који омогућава моделу да приликом обраде сваког токена узме у обзир целокупан контекст улазне секвенце. За разлику од ранијих рекурентних архитектура, које секвенцу обрађују корак по корак, трансформер обрађује све позиције паралелно. Тиме се убрзава обучавање и олакшава рад са дугим текстуалним секвенцама. Основни облик механизма пажње са скалираним скаларним производом дефинише се као:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

где су Q, K и V матрице упита, кључева и вредности, а d_k означава димензију вектора кључева [15]. Скалирање фактором $\sqrt{d_k}$ спречава да скаларни

производи постану превелики у високодимензионим просторима, чиме се избегава да функција softmax уђе у области са веома малим градијентима и тако доприноси стабилнијем обучавању.

Оригинална трансформер архитектура састоји се од два дела: енкодера и декодера. Енкодер обрађује улазну секвенцу и формира контекстуалне репрезентације сваког токена, при чему сваки токен има приступ свим осталим токенима у секвенци. Декодер генерише излазну секвенцу токен по токен, при чему сваки нови токен има приступ само претходно генерисаним токенима, што се постиже применом маске у механизму пажње. У пракси, различити модели користе различите делове ове архитектуре: модели намењени разумевању текста, попут BERT-а, користе само енкодер, генеративни модели попут серије GPT користе само декодер, док модели за превођење или сумаризацију могу користити оба дела.

Овакав ауторегресивни начин генерисања значи да модел у сваком кораку бира наредни токен на основу условне расподеле:

$$P(x_t \mid x_1, x_2, \ldots, x_{t-1}; \theta) \tag{2.2}$$

где x_t представља токен на позицији t, а θ параметре модела. Обучавање се врши максимизацијом логаритма веродостојности над великим корпусом текста, тако да модел научи да за произвољан контекст додели високу вероватноћу токенима који се природно јављају у језику.

Обучавање и прилагођавање

Обучавање великих језичких модела типично се одвија у две фазе. У првој фази, која се назива претходно обучавање (енг. pre-training), модел се обучава на великим количинама необрађеног текста, задатком предвиђања следећег токена. Обим података и број параметара значајно утичу на квалитет научених репрезентација — модел GPT-3, на пример, садржи 175 милијарди параметара и обучен је на корпусу од приближно 570 гигабајта текста [1], док су каснији модели попут GPT-4 [12] додатно унапредили перформансе повећањем обима модела и података.

У другој фази, модел се прилагођава специфичним задацима и очекивањима корисника. Један од најзначајнијих приступа је обучавање уз помоћ повратних информација од људи (енг. Reinforcement Learning from Human Fe-

edback, RLHF), којим се модел усмерава да даје одговоре ближе ономе што корисници очекују.

Други начин прилагођавања је фино подешавање (енг. fine-tuning), при чему се модел додатно обучава на мањем скупу података специфичном за одређени домен. Овај приступ може побољшати перформансе модела у том домену, али има неколико недостатака: захтева припремљен скуп података за обучавање и значајне рачунарске ресурсе, а знање стечено на овај начин временом застарева. Поред тога, фино подешавање не решава проблем навођења извора, јер модел и даље одговара на основу знања садржаног у својим параметрима.

Управо ови недостаци представљају додатну мотивацију за приступ Retrieval-Augmented Generation, код кога модел потребне информације не тражи у сопственим параметрима, већ их добија из спољашњег извора у тренутку генерисања одговора.

Ограничења великих језичких модела

Упркос импресивним способностима, велики језички модели имају неколико ограничења која су релевантна за овај рад. Модели су склони халуцинацијама — генерисању тврдњи које делују убедљиво, али су фактички нетачне [7]. Ово је последица тога што модел функционише на основу статистичких образаца, без могућности да провери исправност онога што генерише. Поред тога, знање модела ограничено је на податке доступне у тренутку обучавања, па модел не може одговорити на питања о догађајима насталим након тог тренутка. Модел такође не може навести изворе из којих потиче информација, што отежава верификацију одговора. Коначно, контекстни прозор модела, иако може бити велик, и даље је ограничен у односу на величину типичних корпуса докумената.

2.2 Генерисање појачано претрагом

Приступ генерисања појачаног претрагом (RAG) комбинује механизам претраге докумената са генеративним језичким моделом. Идеја је да се одговори не формирају искључиво на основу параметарског знања модела, већ да се заснивају на конкретним изворима информација пронађеним у корпусу

докумената. Овај приступ предложен је 2020. године [10] и од тада је постао један од најраспрострањенијих начина за побољшање поузданости великих језичких модела [6].

RAG има неколико практичних предности. Знање система може се ажурирати једноставном заменом или допуном докумената у корпусу, без потребе за поновним обучавањем модела. Претрагом се издвајају само релевантни делови корпуса, чиме се избегава потреба да модел обрађује целокупан корпус одједном. Такође, за сваки генерисани одговор познато је из којих сегмената је формиран, што омогућава проверу и навођење извора.

Фазе RAG система

Типичан RAG систем састоји се од три фазе: индексирања, претраге и генерисања. У фази индексирања, документи из корпуса се припремају за претрагу — деле се на мање сегменте, за сваки сегмент се израчунава векторска репрезентација, а добијени вектори се складиште у векторској бази. Овај корак се извршава само једном, пре почетка рада система.

У фази претраге, на основу постављеног питања проналазе се сегменти који су семантички најсличнији упиту. Питање корисника се пресликава у исти векторски простор као и сегменти, а затим се на основу сличности вектора издваја k најрелевантнијих сегмената.

У фази генерисања, пронађени сегменти се заједно са питањем прослеђују генеративном моделу, који на основу тог контекста формира одговор. Питање и сегменти спајају се у јединствен промпт, који може садржати и додатне инструкције — на пример, да модел одговара искључиво на основу приложеног контекста.

Формално, нека је q питање корисника, $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ скуп сегмената у корпусу, а f функција претраге која враћа k најрелевантнијих сегмената:

$$\mathcal{R}(q) = f(q, \mathcal{D}, k) = \{d_{i_1}, d_{i_2}, \ldots, d_{i_k}\} \tag{2.3}$$

Генеративни модел G затим формира одговор a на основу питања и пронађених сегмената:

$$a = G(q, \mathcal{R}(q)) \tag{2.4}$$

Варијанте RAG архитектуре

RAG системи могу се реализовати на више начина, у зависности од тога како су организоване фазе претраге, допуне контекста и генерисања одговора. Различите варијанте ове архитектуре разликују се пре свега по сложености система и степену интеракције између компоненте за претрагу и генеративног модела [6].

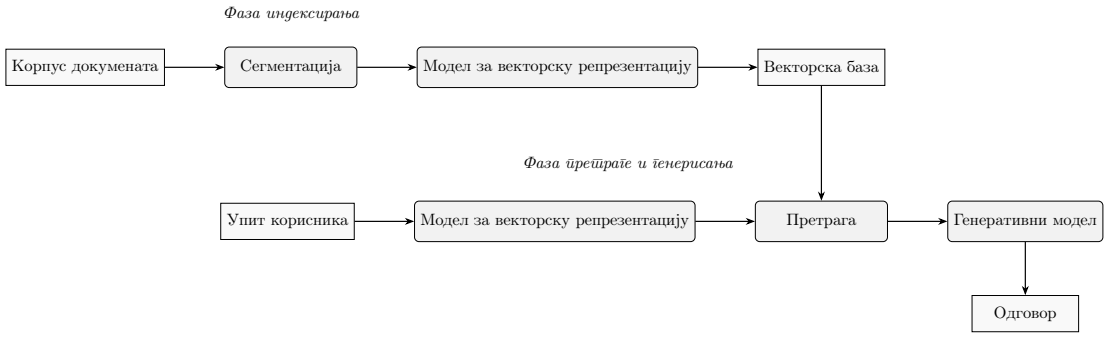
Најједноставнија варијанта позната је као наивни RAG (енг. naive RAG), који прати описани поступак: индексирање, једнократна претрага и генерисање одговора. Овај приступ је једноставан за имплементацију, али има ограничења када иницијална претрага не пронађе довољно релевантне сегменте или када одговор захтева информације из више делова корпуса.

Напредни RAG (енг. advanced RAG) уводи додатне кораке са циљем побољшања квалитета претраге, као што су преформулација упита, ренгирање резултата помоћу додатног модела или коришћење хибридне претраге која комбинује семантичку и лексичку претрагу.

Модуларни RAG (енг. modular RAG) представља најфлексибилнију варијанту, у којој се систем састоји од независних модула који се могу комбиновати на различите начине. Ова архитектура омогућава, између осталог, итеративну претрагу у којој се процес претраге и генерисања понавља више пута.

Систем имплементиран у овом раду може се сврстати у напредни RAG приступ, јер поред основне претраге укључује хибридну претрагу и ренгирање резултата. Са друге стране, систем не користи поступке карактеристичне за модуларни RAG, као што су итеративна претрага или аутоматска преформулација упита.

Општи преглед архитектуре RAG система приказан је на слици 2.1.



Слика 2.1: Општи преглед архитектуре RAG система

Квалитет RAG система зависи од сваке фазе у његовој архитектури. Неадекватна сегментација докумената може довести до мање прецизне претраге, док модел за векторску репрезентацију који не успе да довољно добро представи значење текста може пропустити релевантне сегменте. Са друге стране, чак и када су релевантне информације присутне у контексту, генеративни модел може дати непотпун или недовољно прецизан одговор. Због тога се истраживања у овој области не усмеравају само на крајњи квалитет одговора, већ и на утицај појединачних компоненти система [6]. Ове компоненте, које обухватају сегментацију докумената, векторске репрезентације текста, методе претраге и евалуацију, описане су у наредним одељцима.

2.3 Сегментација докумената

Сегментација докумената је процес поделе текста на мање делове који се у контексту RAG система називају сегментима (енг. chunks). Овај корак је неопходан јер документи у корпусу често садрже више текста него што генеративни модел може да обухвати у оквиру свог контекстног прозора, док претрага на нивоу целих докумената може дати недовољно прецизне резултате [6]. Циљ сегментације је да сваки сегмент буде довољно мали за ефикасну претрагу, али и довољно велик да задржи смислену целину.

Избор стратегије сегментације директно утиче на квалитет претраге и квалитет генерисаних одговора. Сувише кратки сегменти могу изгубити контекст потребан за разумевање садржаја, док сувише дуги сегменти често обухватају више различитих тема, што смањује прецизност претраге [8].

Најједноставнија и најчешће коришћена стратегија је сегментација фиксне

дужине (енг. fixed-size chunking). Документ се дели на сегменте унапред одређене дужине, изражене у токенима или карактерима, при чему се може водити рачуна да се реченице не прекидају на неприродним местима. Суседни сегменти могу се преклапати (енг. overlap), тако да одређен број токена са краја једног сегмента буде укључен и на почетку наредног. На тај начин смањује се ризик од губитка информација на границама сегмената.

Поред сегментације фиксне дужине, постоје и приступи који узимају у обзир структуру документа. Један такав приступ је сегментација на основу секција (енг. section-aware chunking), у којој се границе сегмената усклађују са насловима и поднасловима у документу, тако да сваки сегмент одговара тематски кохерентном делу текста. Сродан приступ је контекстуална сегментација, у којој се сваком сегменту додаје префикс са информацијом о његовом положају у документу, на пример наслов чланка и наслов секције из које сегмент потиче. Овакав префикс моделу за векторску репрезентацију пружа додатне податке о пореклу сегмента, што може побољшати квалитет добијене репрезентације.

Два кључна параметра сваке стратегије сегментације су величина сегмента и величина преклапања. Величина сегмента одређује колико текста сваки сегмент садржи, док величина преклапања одређује колико се текста дели између суседних сегмената. Ови параметри се најчешће изражавају у токенима.

Избор оптималних параметара зависи од карактеристика корпуса, природе питања и модела за векторску репрезентацију текста. Не постоји универзално најбоља стратегија, због чега се у овом раду различите стратегије сегментације пореде експериментално.

2.4 Векторске репрезентације текста

Векторске репрезентације текста (енг. text embeddings) представљају текст као векторе фиксне димензије у високодимензионом векторском простору. Основна идеја је да семантички слични текстови имају блиске векторске репрезентације, док су семантички различити текстови представљени удаљенијим векторима [11]. У RAG системима ове репрезентације омогућавају проналажење релевантних сегмената на основу семантичке сличности са корисничким упитом.

Историјски развој

Рани приступи представљању текста, попут TF-IDF репрезентације, описивали су документе као ретке векторе високе димензије, при чему је свака димензија одговарала једном термину из речника. Иако су овакви приступи једноставни и ефикасни, они не обухватају семантичке односе између речи. На пример, синоними се посматрају као различити термини, док контекст у коме се реч јавља не утиче на њену репрезентацију.

Увођењем векторских репрезентација речи, као што је Word2Vec, омогућено је учење вектора који боље одражавају семантичке односе између речи. Такви модели могу да поставе семантички сродне речи ближе једне другима у векторском простору. Ипак, они свакој речи додељују један фиксан вектор, без обзира на контекст у коме се та реч појављује.

Ово ограничење ублажавају контекстуалне репрезентације, уведене моделима попут BERT-а. Код ових модела иста реч може добити различиту репрезентацију у зависности од реченице или ширег контекста у коме се јавља. Савремени модели за векторску репрезентацију текста на нивоу реченица и пасуса надовезују се на ову идеју, јер настоје да представе значење целог текстуалног сегмента, а не само појединачних речи.

Савремени модели за векторску репрезентацију

Формално, модел за векторску репрезентацију текста може се описати као функција E која текст t пресликава у вектор фиксне димензије d:

$$E(t) \in R^d \tag{2.5}$$

Сличност између два текста t_1 и t_2 затим се може мерити на основу сличности њихових векторских репрезентација. У пракси се могу користити различите мере, као што су скаларни производ, Еуклидско растојање или косинусна сличност. У системима заснованим на векторским репрезентацијама текста најчешће се користи косинусна сличност:

$$\mathrm{sim}(t_1, t_2) = \frac{E(t_1) \cdot E(t_2)}{\|E(t_1)\| \cdot \|E(t_2)\|} \tag{2.6}$$

Вредност косинусне сличности припада интервалу $[-1, 1]$, при чему вредности блиске 1 указују на високу сличност између вектора, а тиме и на већу семантичку сличност текстова.

Савремени модели за векторску репрезентацију текста често се обучавају применом контрастивног учења (енг. contrastive learning) [11]. Модел учи на паровима или групама текстова, при чему семантички сродни текстови, као што су упит и релевантан пасус, чине позитивне примере, док независни текстови чине негативне. Током обучавања, модел се подешава тако да позитивне примере пресликава у блиске тачке у векторском простору, а негативне у удаљеније.

Архитектурално, многи савремени модели за векторску репрезентацију текста заснивају се на трансформер енкодерима. Након обраде улазног текста, репрезентације појединачних токена обједињују се у један вектор који представља значење целе реченице, пасуса или сегмента документа.

Карактеристике модела

Модели за векторску репрезентацију текста разликују се по више карактеристика које утичу на њихову примену у RAG системима. Једна од њих је димензија излазног вектора, која одређује величину векторске репрезентације текста. Већа димензија може омогућити богатије представљање семантичких информација, али истовремено повећава меморијске захтеве и време потребно за израчунавање сличности.

Друга важна карактеристика је максимална дужина улазног текста. Она одређује колико текста модел може да обради у једном пролазу. Текстови који премашују ову границу морају бити скраћени или подељени на мање делове, што може довести до губитка дела контекста. Поред тога, модели се разликују и по језицима које подржавају — неки подржавају само један језик, док други раде са више језика. На пример, модел BGE-M3 подржава више од 100 језика, што га чини погодним за рад са разноврсним корпусима [2].

Избор модела

Комерцијални модели, попут OpenAI породице text-embedding-3 [13], омогућавају једноставну интеграцију и високе перформансе, али захтевају приступ удаљеном сервису и подразумевају трошкове по употреби. Модели отвореног кода, попут BGE-M3 [2], могу се покретати локално, што пружа већу контролу над системом и мању зависност од спољашњих сервиса.

Модел BGE-M3 заслужује посебну пажњу јер подржава векторску репрезентацију текста, ретку лексичку репрезентацију и репрезентацију на нивоу појединачних токена. Векторска репрезентација пресликава цео текст у један вектор, док ретка лексичка репрезентација додељује научене тежине појединачним токенима, што овај модел чини погодним за хибридну претрагу.

Избор модела за векторску репрезентацију директно утиче на квалитет претраге, па поређење различитих модела представља један од кључних експеримената у овом раду.

2.5 Претрага докумената

Претрага докумената је компонента RAG система која на основу постављеног питања проналази сегменте најрелевантније за формирање одговора. Избор приступа претрази значајно утиче на квалитет пронађеног контекста, а тиме и на квалитет генерисаних одговора. У овом одељку описане су три стратегије претраге — Претрага на основу векторских репрезентација, лексичка и хибридна — као и поступак ренгирања резултата.

Претрага на основу векторских репрезентација

Претрага на основу векторских репрезентација (енг. dense retrieval) заснива се на семантичкој сличности између векторских репрезентација упита и сегмената докумената [9]. Упит корисника и сегменти корпуса пресликавају се у исти векторски простор. Затим се проналазе сегменти чије су репрезентације најсличније репрезентацији упита, а сличност се најчешће мери косинусном сличношћу.

У пракси, векторске репрезентације свих сегмената се унапред израчунавају и складиште у векторској бази података. Приликом претраге израчунава се само репрезентација упита, а затим се помоћу алгоритама за приближну претрагу најближих суседа (енг. approximate nearest neighbor, ANN) проналази k сегмената са највећом сличношћу. Овакав приступ омогућава ефикасну претрагу и у великим корпусима.

Предност претраге на основу векторских репрезентација је способност проналажења семантички релевантних сегмената чак и када упит и сегмент не деле исте речи. На пример, упит о "узроцима сукоба" може пронаћи сегмент

који говори о "разлозима избијања рата", иако се кључне речи не поклапају. Недостатак је зависност од квалитета модела за векторску репрезентацију — ако семантичка сличност није довољно висока, претрага може пропустити сегменте који садрже тачне термине или имена из упита.

Лексичка претрага

Лексичка претрага проналази релевантне сегменте на основу поклапања термина између упита и текста сегмента. Најпознатији алгоритам ове врсте је BM25 [14], који сегментима додељује скор на основу тога колико се често термини из упита јављају у сегменту и колико су ти термини информативни у односу на цео корпус. За упит q састављен од термина $q_1, q_2, \ldots, q_n$, BM25 скор сегмента d дефинише се као:

$$\text{BM25}(q, d) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \tag{2.7}$$

где је $f(q_i, d)$ број појављивања термина q_i у сегменту d, $|d|$ дужина сегмента, avgdl просечна дужина сегмената у корпусу, а k_1 и b параметри који контролишу засићење фреквенције и утицај дужине сегмента. Вредност $\text{IDF}(q_i)$ одражава реткост термина q_i у корпусу — термини који се ретко јављају носе више информације и добијају већу тежину.

Предност лексичке претраге је прецизност код упита који садрже тачне термине, имена, датуме или специфичне појмове. Њено главно ограничење је то што не препознаје сегменте који исту информацију изражавају другачијим речима, што је управо снага претраге на основу векторских репрезентација.

Хибридна претрага

Хибридна претрага комбинује претрагу на основу векторских репрезентација и лексичку претрагу како би искористила предности оба приступа [6]. Претрага на основу векторских репрезентација доприноси проналажењу семантички сличних сегмената, док лексичка претрага обезбеђује да сегменти са тачним терминима из упита не буду изостављени.

Да би се резултати две претраге објединили, користи се поступак фузије рангираних листа (енг. Reciprocal Rank Fusion, RRF) [4]. Уместо да се ослања на саме вредности скорова, које код претраге на основу векторских репре-

зентација и лексичке претраге нису директно упоредиве, овај метод узима у обзир само позиције сегмената у појединачним листама. За сегмент d који се у претрази на основу векторских репрезентација налази на позицији $r_1(d)$, а у лексичкој на позицији $r_2(d)$, RRF скор рачуна се као:

$$\mathrm{RRF}(d) = \frac{1}{c + r_1(d)} + \frac{1}{c + r_2(d)} \tag{2.8}$$

где је c константа (у пракси најчешће $c = 60$) која ублажава утицај разлике између високих и ниских позиција. Сегменти који су високо рангирани у обе листе добијају највећи укупни скор.

Ренгирање резултата

Ренгирање (енг. reranking) представља додатни корак који се примењује након почетне претраге, са циљем да се побољша редослед пронађених кандидата. У почетној претрази, сличност између упита и сегмента обично се израчунава на основу њихових одвојено добијених векторских репрезентација. Насупрот томе, ренгирање користи додатни модел који истовремено обрађује пар који чине упит и сегмент, а затим процењује колико је дати сегмент релевантан за упит.

Модели за ренгирање најчешће се заснивају на архитектури унакрсног енкодера (енг. cross-encoder). У овом приступу упит и сегмент се спајају у јединствен улаз, након чега модел генерише скор релевантности. То омогућава прецизније поређење упита и сегмента, јер модел може да анализира њихову међусобну интеракцију, укључујући синтаксичке и семантичке везе које једноставније поређење векторских репрезентација не мора увек да препозна.

Недостатак овог приступа је већа рачунска сложеност. Пошто се сваки пар упита и сегмента мора обрадити посебно, ренгирање је знатно захтевније од почетне претраге. Због тога се овај корак обично не примењује над целим корпусом, већ само над мањим скупом кандидата који је претходно издвојен неком ефикаснијом методом претраге.

Типичан ток хибридне претраге са ренгирањем може се описати кроз неколико корака. Најпре се над корпусом спроводе претрага на основу векторске репрезентације и лексичка претрага, затим се добијене листе резултата комбинују поступком фузије, а на крају се издвојени кандидати ренгирају

помоћу унакрсног енкодера. На тај начин систем комбинује ефикасност почетне претраге са прецизнијом проценом релевантности у завршној фази.

2.6 Евалуација RAG система

Евалуација RAG система обухвата процену две компоненте: квалитета претраге и квалитета генерисаних одговора. За сваку од њих користе се посебне метрике, а поред аутоматских метрика све чешће се примењује и приступ у коме велики језички модел оцењује квалитет одговора [5].

Метрике квалитета претраге

Метрике квалитета претраге користе се за процену тога колико успешно систем проналази релевантне сегменте за дато питање. Њихово израчунавање захтева евалуациони скуп у коме је за свако питање унапред познато који сегменти садрже информације потребне за формирање одговора.

Одзив при првих k резултата (енг. Recall@k) мери колики се део релевантних сегмената налази међу првих k враћених сегмената:

$$\text{Recall@k} = \frac{|\mathcal{R}_k \cap \mathcal{G}|}{|\mathcal{G}|} \tag{2.9}$$

где је $\mathcal{R}_k$ скуп од k враћених сегмената, а $\mathcal{G}$ скуп свих релевантних сегмената за дато питање. Вредност Одзив@k припада интервалу $[0, 1]$, при чему вредност 1 значи да су сви релевантни сегменти пронађени међу k враћених сегмената.

Средња просечна прецизност (енг. Mean Average Precision, MAP) узима у обзир не само да ли су релевантни сегменти пронађени, већ и на којим позицијама се налазе у ранг-листи. За једно питање, просечна прецизност се рачуна као:

$$\text{AP} = \frac{1}{|\mathcal{G}|} \sum_{j=1}^{k} \text{P}(j) \cdot \text{rel}(j) \tag{2.10}$$

где је $\text{P}(j)$ прецизност на позицији j, а $\text{rel}(j)$ индикаторска функција која има вредност 1 ако је сегмент на позицији j релевантан, а 0 у супротном. Метрика MAP се добија као просек вредности AP преко свих питања у евалуационом

скупу. У односу на Одзив@k, ова метрика је строжа јер узима у обзир и редослед резултата, па више вреднује системе код којих се релевантни сегменти налазе ближе врху ранг-листе.

Метрике квалитета генерисаних одговора

Поред квалитета претраге, потребно је проценити и квалитет самих генерисаних одговора. Одговор се може посматрати кроз више димензија, а у овом раду разматрају се четири: тачност, утемељеност, релевантност и потпуност. Тачност изражава да ли су тврдње у одговору чињенично исправне. Утемељеност процењује да ли су тврдње поткрепљене пронађеним контекстом, односно да ли модел износи само оно што је присутно у датим сегментима. Релевантност изражава степен у ком се одговор односи на постављено питање, док потпуност процењује да ли одговор обухвата све кључне информације потребне за дато питање.

Ове димензије су међусобно независне. На пример, одговор може бити тачан и утемељен у контексту, али непотпун ако изоставља део релевантних информација. Слично, одговор може бити потпун и релевантан, али нетачан ако садржи халуцинације.

Приступ LLM као судија

Мануелна евалуација генерисаних одговора је поуздана, али временски захтевна и тешко скалабилна за велике евалуационе скупове. Алтернативни приступ је LLM као судија, у коме велики језички модел оцењује квалитет генерисаних одговора на основу унапред дефинисаних критеријума [16].

У овом приступу, моделу се прослеђују питање, генерисани одговор, пронађени сегменти и, ако постоји, референтни одговор, на основу чега модел додељује нумеричке оцене по свакој димензији квалитета и образлаже своје одлуке. Предности оваквог приступа су скалабилност и конзистентност — исти критеријуми примењују се на све одговоре, без варијација које су карактеристичне за мануелну евалуацију.

Међутим, приступ LLM као судија има и ограничења. Модел који оцењује може имати систематске пристрасности, на пример склоност ка дужим одговорима или одговорима генерисаним сличним моделом. Поред тога, тачност оцена може зависити од сложености питања и од прецизности дефинисаних

критеријума [16]. Из тих разлога, поређење аутоматских оцена са мануелном евалуацијом представља важан корак у валидацији овог приступа, што је један од предмета анализе у овом раду.

Глава 3

Методологија

3.1 Опис корпуса

За потребе експерименталне анализе формиран је корпус историјских чланака преузетих са Википедије. Историјски домен је изабран зато што садржи велики број чињеничних података, временских односа, личности, догађаја и узрочно-последичних веза. Поред тога, историјски текстови често садрже информације распоређене кроз више делова документа, што их чини погодним за испитивање система који треба да пронађу и повежу више релевантних извора информација приликом генерисања одговора. На тај начин омогућена је евалуација система како на једноставним чињеничним питањима, тако и на питањима која захтевају интеграцију информација из више сегмената.

Корпус је прикупљен у два корака. У првом кораку коришћен је модул за преузимање података, који је путем Википедијиног програмског интерфејса преузимао чланке из унапред изабраних историјских категорија. У другом кораку примењен је поступак филтрирања корпуса, са циљем уклањања нерелевантних или недовољно квалитетних чланака. Филтрирање је обухватило уклањање појединих категорија, чланака чији наслови указују на листе, библиографије, временске линије, индексе или друге помоћне странице, као и текстова који не задовољавају минималне критеријуме дужине и тематске релевантности.

Критеријуми филтрирања били су дефинисани тако да задрже чланке који садрже довољно историјског садржаја и који су погодни за формирање сегмената над којима се касније врши претрага. Поред минималне дужине текста, коришћена је и провера присуства кључних историјских појмова, као што су

појмови повезани са ратовима, биткама, царствима, владарима, династијама, војним походима и споразумима.

Коначни корпус садржи 326 чланака распоређених у 13 историјских категорија. Категорије покривају широк временски распон, од античке историје до модерног доба, и обухватају теме као што су Први светски рат, Крсташки ратови, Наполеонови ратови, историја Европе, Америчка револуција, стари Египат, Ахеменидско царство, Османско царство, стари Рим, Други светски рат, стара историја, Амерички грађански рат и стара Грчка. Расподела чланака по категоријама није потпуно уједначена: најзаступљенија је категорија Први светски рат са 47 чланака, док је најмање заступљена категорија стара Грчка са 10 чланака. Циљ формирања корпуса није био равномерна заступљеност свих категорија, већ добијање тематски разноврсног, али доменски ограниченог скупа докумената погодног за евалуацију RAG система.

Просечна дужина чланка у корпусу износи приближно 18.700 карактера. Најкраћи чланак има нешто више од 3.200 карактера, док најдужи чланак има око 133.000 карактера. Медијална дужина чланка износи приближно 10.500 карактера, што указује на то да корпус садржи већи број краћих и средње дугих чланака, али и мањи број веома дугачких текстова. Ова разлика у дужини чланака важна је за даљу обраду, јер директно утиче на број сегмената који се добијају након поделе докумената.

Сваки документ у корпусу представља један Википедијин чланак и садржи текст чланка, наслов, адресу странице, дужину текста и изворну категорију. Ови метаподаци омогућавају праћење порекла докумената и касније су коришћени приликом сегментације, изградње векторске базе и анализе резултата претраге.

3.2 Стратегије сегментације

У раду су испитане три стратегије сегментације докумената, које се разликују по величини сегмената, величини преклапања и начину на који се структура документа узима у обзир.

Прва стратегија, означена као базна, користи сегменте величине 512 токена са преклапањем од 50 токена. Токенизација је извршена помоћу библиотеке tiktoken, коришћењем токенизатора модела GPT-4o-mini, који је уједно и генеративни модел коришћен у експериментима. Овакав избор обезбеђује да

дужина сегмената буде конзистентна са начином на који генеративни модел обрађује текст. Приликом поделе текста поштоване су границе реченица, тако да ниједан сегмент не почиње нити се завршава на средини реченице.

Друга стратегија, означена као стратегија малих сегмената, користи сегменте величине 256 токена са преклапањем од 25 токена. Ова стратегија је испитана како би се утврдило да ли мањи сегменти омогућавају прецизнију претрагу релевантног контекста, уз могући трошак у виду губитка ширег контекста садржаног у већим сегментима.

Трећа стратегија је контекстуална сегментација, која користи структуру документа приликом формирања сегмената. У овој стратегији текст се најпре дели на секције на основу наслова присутних у документу, док се библиографске секције, као што су референце, спољашње везе и додатна литература, изостављају из даље обраде. Сваком добијеном сегменту додаје се контекстуални префикс у формату "[Наслов чланка > Наслов секције]". Тај префикс постаје део текста који се користи за израчунавање векторске репрезентације, чиме се моделу пружа додатна информација о пореклу и тематском контексту сегмента.

Уколико је текст секције дужи од максималне дозвољене величине сегмента, секција се додатно дели на мање делове уз преклапање суседних сегмената. Пошто се контекстуални префикс додаје сваком сегменту и улази у укупан број токена, дужина самог садржаја сегмента зависи од дужине наслова чланка и наслова секције. Због тога сегменти добијени овом стратегијом не морају имати једнаку дужину, али сваки задржава информацију о секцији из које потиче.

3.3 Формирање евалуационог скупа

За потребе евалуације система формиран је скуп од 163 питања. Питања су генерисана аутоматски помоћу модела GPT-4o и формулисана су на енглеском језику, при чему су као основа коришћени насумично изабрани сегменти из базне сегментације, бирани стратификовано по историјским категоријама како би све категорије биле заступљене. За свако питање, као референтни релевантни сегменти (енг. ground truth) означени су сегменти на основу којих је питање генерисано.

Након аутоматског генерисања, сва питања су мануелно прегледана. Из

скупа су уклоњена питања која нису била довољно јасна, питања на која се могло одговорити без увида у сегмент, као и питања чији изворни сегменти нису садржали довољно информација за потпун одговор. Овим кораком обезбеђено је да свако питање у коначном скупу има јасно дефинисан и проверен скуп релевантних сегмената.

Питања су подељена у три категорије према начину на који се долази до одговора. Прву категорију чине питања која захтевају један релевантан сегмент (енг. single-chunk), укупно 57 питања. Ова питања односе се на чињеницу садржану у једном сегменту и захтевају директно проналажење одговарајућег дела текста.

Другу категорију чине питања која захтевају више релевантних сегмената (енг. multi-chunk), укупно 54 питања. Код ових питања два релевантна сегмента потичу из истог чланка, а одговор зависи од повезивања информација из оба сегмента, на пример кроз поређење или успостављање узрочно-последичне везе. Овај тип питања представља већи изазов за компоненту претраге, јер је потребно пронаћи оба релевантна сегмента, а не само један.

Трећу категорију чине сложена питања (енг. complex), укупно 52 питања. Иако се заснивају на једном релевантном сегменту, ова питања не захтевају пуко проналажење чињенице, већ тумачење, сумирање или извођење закључка на основу садржаја сегмента. На тај начин се испитује способност система да на основу пронађеног контекста генерише одговор који захтева дубље разумевање, а не само преписивање информације.

Овако формиран евалуациони скуп омогућава анализу перформанси система не само на укупном нивоу, већ и у зависности од типа и сложености постављеног питања.

3.4 Модели коришћени у систему

Експериментални систем користи неколико језичких модела, од којих сваки има посебну улогу: моделе за векторску репрезентацију текста, генеративни модел за формирање одговора и модел за аутоматску процену квалитета генерисаних одговора.

Модели за векторску репрезентацију

За претрагу на основу векторских репрезентација испитана су три модела за векторску репрезентацију текста. Прва два, text-embedding-3-small и text-embedding-3-large, припадају комерцијалној породици модела компаније OpenAI [13] и доступни су преко програмског интерфејса. Модел text-embedding-3-small генерише векторе димензије 1536, док text-embedding-3-large генерише векторе димензије 3072, чиме може представити финије семантичке разлике, али уз веће меморијске и рачунарске захтеве.

Трећи модел, BGE-M3 [2], је модел отвореног кода који генерише векторе димензије 1024 и може се покретати локално. Поред векторске репрезентације текста, овај модел подржава и ретку лексичку репрезентацију, што га чини погодним за хибридну претрагу. У оквиру хибридне претраге, лексичка компонента реализована је помоћу алгоритма BM25, а за реренгирање кандидата коришћен је модел BGE-reranker-v2-m3, заснован на архитектури унакрсног енкодера.

Генеративни модел

За генерисање одговора у свим конфигурацијама коришћен је модел GPT-4o-mini компаније OpenAI, са температуром постављеном на 0, чиме се обезбеђује детерминистичко генерисање. Овај модел изабран је као компромис између квалитета одговора и трошкова, с обзиром на велики број конфигурација и питања у евалуационом скупу. Исти генеративни модел коришћен је у свим конфигурацијама како би се обезбедило да разлике у квалитету одговора потичу од компоненти претраге, а не од промене генеративног модела.

Модел за аутоматску евалуацију

За аутоматску евалуацију квалитета генерисаних одговора применом приступа LLM као судија коришћен је модел Claude Opus 4.6 компаније Anthropic. У овом приступу велики језички модел процењује одговоре на основу унапред дефинисаних критеријума. Пошто је за генерисање одговора коришћен модел компаније OpenAI, за евалуацију је изабран модел из друге породице модела, чиме се смањује ризик од систематске пристрасности у оцењивању.

Окружење и алати

Сви експерименти имплементирани су у програмском језику Python. За токенизацију при сегментацији коришћена је библиотека tiktoken, која користи исти токенизатор као модели компаније OpenAI. За складиштење и претрагу векторских репрезентација коришћена је векторска база ChromaDB [3], која подржава претрагу на основу косинусне сличности. Модели отвореног кода BGE-M3 и BGE-reranker-v2-m3 покретани су локално, на изнајмљеном серверу RunPod са графичком картицом NVIDIA RTX 4090 и 24 гигабајта видео-меморије.

3.5 Конфигурације и ток експеримената

Поређење различитих компоненти RAG система спроведено је по принципу контролисане варијације: у сваком кораку мења се само једна компонента система, док остале остају непромењене. Овакав приступ омогућава да се утицај сваке компоненте сагледа изоловано, без мешања са ефектима других промена.

Полазна тачка је базна конфигурација, која користи сегментацију 512/50, модел text-embedding-3-small, претрагу на основу векторских репрезентација и $k = 5$ враћених сегмената. Ова конфигурација представља једноставан RAG систем и служи као референтна тачка за поређење.

Ток експеримената организован је тако да се најбоља конфигурација из једног корака користи као полазиште за наредни. Најпре је испитан утицај стратегије сегментације, поређењем базне, мале и контекстуалне сегментације. Затим је варирана вредност броја враћених сегмената ($k \in \{3, 5, 10\}$). У наредном кораку поређена су три модела за векторску репрезентацију текста, а на крају је претрага на основу векторских репрезентација упоређена са хибридном претрагом, која комбинује BGE-M3, BM25, спајање рангираних листа поступком RRF и реренгирање.

Укупно је испитано осам конфигурација, приказаних у табели 3.1. Свака конфигурација евалуирана је на истом скупу од 163 питања, чиме је обезбеђена упоредивост резултата.

Табела 3.1: Преглед свих испитаних конфигурација (Векторска = претрага на основу векторских репрезентација)

Бр.	Модел за репрезентацију	k	Сегментација	Претрага
1	text-embedding-3-small	5	512/50	векторска
2	text-embedding-3-small	5	256/25	векторска
3	text-embedding-3-small	5	контекстуална	векторска
4	text-embedding-3-small	3	512/50	векторска
5	text-embedding-3-small	10	512/50	векторска
6	text-embedding-3-large	10	512/50	векторска
7	BGE-M3	10	512/50	векторска
8	BGE-M3	10	512/50	хибридна

За сваки упит, систем обрађује питање кроз исти редослед корака: упит се пресликава у векторску репрезентацију, претражује се векторска база, по потреби се примењује хибридна претрага са ренгирањем, а затим се изабрани сегменти заједно са питањем прослеђују генеративном моделу. Промпт је конструисан тако да модел одговара искључиво на основу приложеног контекста, без ослањања на сопствено параметарско знање, како би квалитет одговора зависио претежно од квалитета претраге. Овај ток приказан је на слици 3.1.

Слика 3.1: Ток обраде упита у експерименталном систему

3.6 Мапирање релевантних сегмената

Питања у евалуационом скупу везана су за сегменте из базне сегментације (512/50), јер су управо на основу тих сегмената и формулисана. Проблем настаје када се испитује нека друга стратегија сегментације: тада су границе сегмената другачије, па релевантни сегменти означени у базној сегментацији више немају директан пандан у новој подели текста. Да би се и за такве стратегије могле израчунати метрике претраге, потребно је пронаћи који сегменти у новој сегментацији одговарају изворним релевантним сегментима.

Ово се решава поређењем на нивоу токена. За сваки релевантан сегмент

из базне сегментације проверава се колико се његов садржај преклапа са сегментима из нове сегментације који потичу из истог чланка. Уколико удео заједничких токена пређе праг од 0,52, такав сегмент се означава као релевантан. Праг је одређен емпиријски — испробавањем неколико вредности изабрана је она која постиже најбољу равнотежу између обухватања свих релевантних сегмената и избегавања погрешних поклапања. Поређење се намерно ограничава на исти чланак, будући да релевантан сегмент увек потиче из конкретног документа, па нема смисла тражити подударања у другим деловима корпуса.

Овакав поступак има и одређено ограничење. Када се један сегмент из базне сегментације у новој подели распадне на више мањих сегмената, сваки од њих који довољно личи на оригинал биће означен као релевантан. Због тога се број релевантних сегмената по питању може разликовати од оног у базној сегментацији, што је узето у обзир при тумачењу резултата, нарочито код стратегије са ситнијим сегментима.

3.7 Метрике евалуације

За евалуацију система коришћене су две групе метрика, дефинисане у одељку 2.6. Квалитет претраге процењен је метрикама Одзив@k и MAP, израчунатим на основу скупа референтних релевантних сегмената из евалуационог скупа.

Квалитет генерисаних одговора оцењен је кроз четири димензије: тачност, утемељеност, релевантност и потпуност. Свака димензија оцењује се на скали од 1 до 5, где оцена 1 означава најнижи, а оцена 5 највиши квалитет. У оквиру приступа LLM као судија, моделу који оцењује задати су прецизни критеријуми за сваку димензију, са описом значења појединачних оцена на скали. Тачност показује у којој мери је одговор у складу са референтним одговором и очекиваним кључним чињеницама. Утемељеност се односи на повезаност одговора са пронађеним контекстом: одговор је боље утемељен ако се његове тврдње могу поткрепити информацијама из враћених сегмената. Релевантност показује колико се одговор непосредно односи на постављено питање, без одступања ка споредним или непотребним информацијама. Потпуност се процењује на основу тога колико су у одговору обухваћене кључне чињенице потребне за дато питање.

Поред нумеричких оцена, за сваки одговор одређује се и доминантан тип грешке, уколико грешка постоји. Разликују се следећи типови: изостављање релевантних информација присутних у контексту, халуцинације (тврдње које контекст не подржава), погрешно приписивање информација изворима који нису међу пронађеним сегментима, и неуспешна претрага, када релевантне информације уопште нису доспеле у контекст. Уколико одговор не садржи значајне грешке, означава се као одговор без грешке.

3.8 Мануелна евалуација и LLM као судија приступ

Поред аутоматске евалуације применом приступа LLM као судија, на подскупу питања спроведена је и мануелна евалуација. Њен циљ је да се провери у којој мери се аутоматске оцене поклапају са оценама које даје човек, као и да се уоче евентуалне систематске разлике између та два приступа.

За мануелну евалуацију изабрано је 45 питања, по 15 из сваке од три раније наведене категорије. За свако питање мануелно су оцењене све четири димензије квалитета, на истој скали од 1 до 5 и по истим критеријумима који су коришћени у аутоматској евалуацији, чиме је обезбеђена упоредивост оцена.

Поређење је извршено анализом просечних одступања између аутоматских и мануелних оцена, и то за сваку димензију квалитета и за сваку категорију питања. Резултати овог поређења приказани су у поглављу са експерименталним резултатима.

Глава 4

Експериментални резултати

У овом поглављу приказани су резултати експеримената спроведених над осам конфигурација RAG система. Резултати су организовани тако да најпре прикажу понашање базне конфигурације, а затим утицај појединачних компоненти система на квалитет претраге и генерисаних одговора. У наставку се резултати анализирају и према типовима питања, типичним грешкама, као и кроз поређење мануелне евалуације са оценама добијеним применом приступа LLM као судија.

4.1 Резултати базне конфигурације

Базна конфигурација користи сегментацију 512/50, модел text-embedding-3-small за векторску репрезентацију, претрагу на основу векторских репрезентација и вредност параметра $k = 5$. Ова конфигурација представља полазну тачку за наредне експерименте и служи као основа за поређење резултата добијених изменом појединачних компоненти система.

На целом евалуационом скупу од 163 питања, базна конфигурација постиже Одзив@5 од 0,868 и MAP од 0,690. Систем тако у већини случајева успешно проналази релевантне сегменте, али вредност MAP указује на то да постоји простор за боље рангирање пронађеног контекста. Када је реч о квалитету генерисаних одговора, просечне оцене износе 4,34 за тачност, 4,72 за утемељеност, 4,67 за релевантност и 4,04 за потпуност, уз просечну латенцију од 8,16 секунди по питању.

Ови почетни резултати показују да и једноставна конфигурација RAG система може да постигне солидан квалитет претраге и одговора. Ипак,

потпуност је најниже оцењена димензија, што значи да одговори понекад не обухватају све очекиване информације. Један од разлога може бити непотпун пронађени контекст — када део релевантних сегмената није међу враћеним резултатима, генеративном моделу недостају информације потребне за потпун одговор.

4.2 Утицај стратегије сегментације

У овом експерименту поређене су три стратегије сегментације: базна сегментација 512/50, стратегија мањих сегмената 256/25 и контекстуална сегментација. Остале компоненте система задржане су као у базној конфигурацији.

Табела 4.1: Поређење стратегија сегментације (Утем. = утемељеност, Рел. = релевантност, Потп. = потпуност)

Стратегија	Одзив@5	MAP	Тачност	Утем.	Рел.	Потп.
Базна (512/50)	0,868	0,690	4,34	4,72	4,67	4,04
Мали сегменти (256/25)	0,695	0,542	4,20	4,69	4,64	3,78
Контекстуална	0,776	0,612	4,25	4,84	4,61	3,89

Базна стратегија сегментације постиже најбоље метрике претраге међу посматраним стратегијама. У односу на њу, стратегија мањих сегмената доводи до приметног пада вредности Одзив@5 и MAP. Овакав резултат указује на то да краћи сегменти, иако могу бити прецизнији у ужем локалном контексту, чешће не садрже довољно информација потребних за поуздано препознавање релевантног садржаја. Контекстуална сегментација остварује боље резултате од стратегије мањих сегмената, али и даље заостаје за базном сегментацијом у погледу метрика претраге.

Разлике у квалитету генерисаних одговора мање су изражене него разлике у метрикама претраге. Посебно се издваја контекстуална сегментација, која постиже највишу утемељеност међу посматраним стратегијама, са просечном оценом 4,84. Овај резултат сугерише да додавање информације о наслову чланка и секције може помоћи моделу да одговор боље ослони на приложени контекст. Ипак, базна сегментација постиже најбољи укупан однос између квалитета претраге и квалитета генерисаних одговора, па је задржана као основа за наредне експерименте.

4.3 Утицај броја враћених сегмената

Број враћених сегмената одређује колико контекста добија генеративни модел, па је његов утицај испитан поређењем вредности $k \in \{3, 5, 10\}$. Остале компоненте задржане су као у базној конфигурацији.

Табела 4.2: Утицај броја враћених сегмената

k	Одзив@k	MAP	Тачност	Утем.	Рел.	Потп.
3	0,782	0,662	4,23	4,82	4,59	3,84
5	0,868	0,690	4,34	4,72	4,67	4,04
10	0,933	0,700	4,47	4,76	4,84	4,19

Повећање броја враћених сегмената са три на десет доводи до осетног раста Одзив@k (са 0,782 на 0,933), а уз њега расту и тачност, релевантност и потпуност одговора. Утемељеност, међутим, показује супротан тренд — највиша је при $k = 3$ (4,82), а благо опада како се број сегмената повећава. Ово је очекивано: већи број враћених сегмената повећава шансу да релевантни сегменти буду пронађени, али уједно уноси и више нерелевантног контекста, који модел понекад искористи за тврдње које најрелевантнији сегменти не подржавају у потпуности.

4.4 Утицај модела за векторску репрезентацију

Следећи експеримент бави се утицајем модела за векторску репрезентацију текста. Поређена су три модела — text-embedding-3-small, text-embedding-3-large и BGE-M3 — уз остале компоненте задржане као у базној конфигурацији и $k = 10$ враћених сегмената.

Табела 4.3: Поређење модела за векторску репрезентацију текста

Модел	Одзив@10	MAP	Тачност	Утем.	Рел.	Потп.
text-emb-3-small	0,933	0,700	4,47	4,76	4,84	4,19
text-emb-3-large	0,945	0,717	4,44	4,74	4,79	4,17
BGE-M3	0,945	0,744	4,47	4,78	4,81	4,19

Модели text-embedding-3-large и BGE-M3 постижу исти Одзив@10 (0,945), нешто виши него text-embedding-3-small (0,933). Кључна разлика је у рангирању: BGE-M3 остварује највиши MAP (0,744), што значи да релевантне сегменте смешта на више позиције у листи од друга два модела. Разлике у квалитету одговора при том остају мале, што поново показује да побољшање метрика претраге не повлачи увек сразмерно бољи одговор.

На основу ових резултата, BGE-M3 је изабран као модел за векторску репрезентацију у експерименту са хибридном претрагом.

4.5 Поређење претраге на основу векторских репрезентација и хибридне претраге

Последњи експеримент пореди претрагу на основу векторских репрезентација текста добијених моделом BGE-M3 са хибридном претрагом. Хибридна претрага користи исти модел за добијање векторских репрезентација, али га допуњује лексичком претрагом помоћу BM25, спајањем рангираних листа поступком RRF и реренгирањем помоћу модела BGE-reranker-v2-m3. Обе конфигурације користе базну сегментацију и $k = 10$ враћених сегмената.

Табела 4.4: Поређење векторске и хибридне претраге

Претрага	Одзив@10	MAP	Тачност	Утем.	Рел.	Потп.
Векторска (BGE-M3)	0,945	0,744	4,47	4,78	4,81	4,19
Хибридна	0,957	0,763	4,53	4,77	4,91	4,26

Хибридна претрага постиже најбоље вредности Одзив@10 (0,957) и MAP (0,763) од свих испитаних конфигурација. Раст метрика претраге прати и раст квалитета одговора — тачност се повећава са 4,47 на 4,53, а релевантност са 4,81 на 4,91. Утемељеност остаје готово непромењена (4,77 наспрам 4,78), што значи да додатни контекст из лексичке претраге и реренгирања не нарушава везаност одговора за приложене сегменте.

Цена овог побољшања је дуже кашњење — хибридна претрага троши у просеку 10,04 секунде по питању, наспрам 8,18 секунди код претраге на основу векторских репрезентација, услед додатних корака обраде. Та разлика може бити значајна у применама где је брзина одговора важна.

На крају, у табели 4.5 дато је укупно рангирање свих осам конфигурација, уз метрике претраге, оцене квалитета одговора, просечну латенцију и расподелу типова грешака.

Табела 4.5: Укупно рангирање свих конфигурација (Кашњ. = просечно кашњење по питању)

Ранг	Конфигурација	Одзив@к	MAP	Тачност	Утем.	Рел.	Пот.	Кашњ.
1	Хибридна (BGE-M3)	0,957	0,763	4,53	4,77	4,91	4,26	10,04s
2	Векторска (BGE-M3, k=10)	0,945	0,744	4,47	4,78	4,81	4,19	8,18s
3	Векторска (text-emb-3-small, k=10)	0,933	0,700	4,47	4,76	4,84	4,19	9,70s
4	Векторска (text-emb-3-large, k=10)	0,945	0,717	4,44	4,74	4,79	4,17	8,60s
5	Базна	0,868	0,690	4,34	4,72	4,67	4,04	8,16s
6	Контекстуална сегментација	0,776	0,612	4,25	4,84	4,61	3,89	8,70s
7	Векторска (text-emb-3-small, k=3)	0,782	0,662	4,23	4,82	4,59	3,84	8,70s
8	Мали сегменти	0,695	0,542	4,20	4,69	4,64	3,78	7,80s

Хибридна конфигурација заузима прво место, са најбољим вредностима за Одзив@k, MAP, тачност, релевантност и потпуност. Друго место заузима претрага на основу векторских репрезентација са моделом BGE-M3, која постиже врло сличне оцене квалитета одговора уз приметно ниже кашњење, па представља добар практичан избор када је брзина важна. Занимљиво је да модел text-embedding-3-small заузима виши ранг од модела text-embedding-3-large, захваљујући бољој тачности и релевантности упркос нешто нижим метрикама претраге. Конфигурације при дну табеле издвајају се по знатно већем броју грешака, пре свега изостављања информација и неуспешне претраге.

4.6 Анализа резултата по типовима питања

У овом одељку приказани су резултати најбоље конфигурације (хибридна претрага) раздвојени по типовима питања, чиме се анализира како сложеност питања утиче на квалитет система.

У табели 4.6 приказани су резултати хибридне конфигурације за сваки тип питања.

Разлике између типова питања су изражене. Код питања која захтевају један сегмент, хибридна конфигурација постиже савршен Одзив@10 од 1,000 и MAP од 0,940, уз просечну тачност од 4,96. Код питања која захтевају више сегмената, Одзив@10 опада на 0,907, а MAP на 0,622. Потпуност одговора је најнижа за овај тип питања (3,74), што указује на то да систем често не успева

Табела 4.6: Резултати хибридне конфигурације по типовима питања

Тип питања	n	Одзив@10	MAP	Тачност	Утем.	Потп.
Један сегмент	57	1,000	0,940	4,96	4,98	4,81
Два сегмента	54	0,907	0,622	4,09	4,54	3,74
Сложеније питање	52	0,962	0,714	4,52	4,77	4,19

да искористи све потребне информације чак и када су релевантни сегменти присутни међу враћеним резултатима.

Сложена питања показују занимљив образац: Одзив@10 је висок (0,962), али је потпуност (4,19) нижа од очекиване, што сугерише да је изазов код ових питања више у способности генеративног модела да повеже информације него у самој претрази.

4.7 Анализа типова грешака

Поред нумеричких оцена, за сваки одговор одређен је и доминантан тип грешке. У табели 4.7 приказана је расподела типова грешака за свих осам конфигурација, на целом евалуационом скупу од 163 питања.

Табела 4.7: Расподела типова грешака по конфигурацијама (Без гр. = без грешке, Изост. = изостављање информација, Неусп. = неуспешна претрага, Халуц. = халуцинација, Погр. = погрешно приписивање)

Конфигурација	Без гр.	Изост.	Неусп.	Халуц.	Погр.
Хибридна (BGE-M3)	92	63	2	5	1
Векторска (BGE-M3)	92	64	3	2	2
Векторска (text-emb-3-small, k=10)	96	63	2	1	1
Векторска (text-emb-3-large, k=10)	95	65	2	0	1
Базна	93	60	9	1	0
Контекстуална сегментација	77	72	14	0	0
Векторска (text-emb-3-small, k=3)	76	68	18	1	0
Мали сегменти	70	76	12	4	1

У свим конфигурацијама најчешћи тип грешке је изостављање релевантних информација. Та грешка је доминантна и јавља се у значајном делу одговора, што се поклапа са раније уоченим податком да је потпуност најслабија димензија квалитета. Неуспешна претрага знатно је ређа код конфигурација са $k = 10$ него код оних са мањим бројем враћених сегмената — на

пример, базна конфигурација са $k = 3$ има 18 случајева неуспешне претраге, док их хибридна конфигурација има само 2. Ово је у складу са резултатима из одељка 4.3, где већи број враћених сегмената повећава вероватноћу да релевантни сегменти буду пронађени.

Халуцинације су ретке у свим конфигурацијама, што указује на то да промпт који модел ограничава на приложени контекст успешно сузбија овај тип грешке. Ипак, хибридна конфигурација бележи нешто више халуцинација (5) него претрага на основу векторских репрезентација са истим моделом (2), што може бити последица тога што хибридна претрага доноси разноврснији контекст, укључујући лексички сличне сегменте који понекад наведу модел на нетачан закључак.

4.8 Поређење мануелне евалуације и LLM као судија приступа

Мануелна евалуација спроведена је на подскупу од 45 питања, по 15 из сваке категорије, за свих осам конфигурација. У табели 4.8 приказана су просечна одступања (Δ) између аутоматских и мануелних оцена, израчуната као разлика аутоматске и мануелне оцене (позитивна вредност значи да је аутоматска оцена виша).

Табела 4.8: Просечна одступања између аутоматских и мануелних оцена (ΔТачн. = разлика у тачности, ΔУтем. = разлика у утемељености, ΔРел. = разлика у релевантности, ΔПотп. = разлика у потпуности)

Конфигурација	ΔТачн.	ΔУтем.	ΔРел.	ΔПотп.
Векторска (text-emb-3-small, k=10)	−0,22	−0,04	−0,02	−0,36
Векорска (text-emb-3-large, k=10)	−0,20	+0,11	−0,02	−0,22
Векторска (BGE-M3)	−0,11	+0,15	+0,03	+0,06
Хибридна	+0,09	−0,08	+0,25	−0,16
Базна	0,00	+0,31	+0,04	+0,16
Векторска (text-emb-3-small, k=3)	−0,16	+0,20	0,00	−0,16
Контекстуална сегментација	−0,20	+0,40	+0,04	−0,24
Мали сегменти	+0,09	+0,20	−0,02	−0,09

Одступања се углавном крећу у опсегу $\pm$0,3, што говори о доброј укупној саглашености аутоматске и мануелне евалуације. Ипак, примећују се две

супротне тенденције. Код потпуности, аутоматска оцена је у просеку нешто нижа од људске (за око $0,13$), што значи да модел строже процењује да ли је одговор обухватио све тражене информације. Код утемељености је обрнуто — аутоматска оцена је нешто виша (за око $0,15$): пошто пред собом има цео враћени контекст, модел лакше препознаје да је нека тврдња њиме поткрепљена, док је човек у тој процени по правилу опрезнији.

Највећа неслагања јављају се код питања која захтевају више сегмената. Тако код базне конфигурације са $k = 10$ разлика у потпуности за овај тип питања достиже $-0,53$, а у тачности $-0,47$. Аутоматска и мануелна евалуација се, дакле, највише разилазе управо код одговора који захтевају повезивање информација из више извора, што је и методолошки захтевнији случај за процену квалитета.

Глава 5

Дискусија

У овом поглављу тумаче се главни резултати, разматрају разлике између конфигурација и сагледавају ограничења експерименталне поставке.

5.1 Тумачење добијених резултата

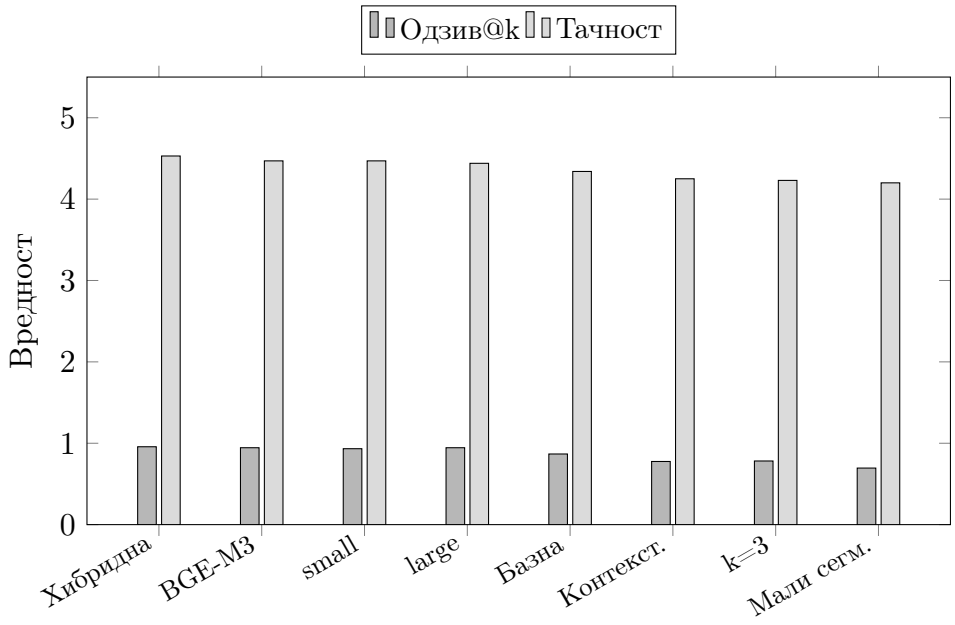
Резултати показују да све испитане компоненте доприносе укупном квалитету система, али и да њихово дејство није увек једноставно. Иако побољшање сваке компоненте начелно подиже квалитет, поједини налази откривају да се добици у претрази и добици у квалитету одговора не крећу увек упоредо.

Најјаснији пример је утицај броја враћених сегмената. Већи број сегмената подиже метрике претраге и већину оцена квалитета, али истовремено благо снижава утемељеност — заједно са релевантним сегментима, у контекст доспева и део оних који нису важни за питање, па се модел повремено на њих ослони при формирању одговора. Тиме се наговештава да више пронађеног контекста није нужно и боље, што отвара питање односа између квалитета претраге и квалитета одговора, коме је посвећен наредни одељак.

5.2 Однос квалитета претраге и квалитета одговора

Слика 5.1 обједињује све конфигурације и открива образац који из појединачних табела није очигледан: Одзив@k варира у широком распону, док тачност одговора свих конфигурација остаје унутар уског појаса. Другим ре-

чима, велике разлике у квалитету претраге преводе се у тек мале разлике у квалитету одговора.



Слика 5.1: Поређење конфигурација по Одзив@k и тачности генерисаних одговора

Овај несклад је један од кључних налаза рада и потврђује се на више места. Контекстуална сегментација, упркос осетно слабијој претрази, постиже највишу утемељеност међу свим конфигурацијама. Модел text-embedding-3-large има боље метрике претраге од мањег модела, али готово исти квалитет одговора. У оба случаја јача претрага не доноси сразмерно бољи одговор.

Овакво понашање може се објаснити на неколико начина. Генеративни модел има ограничену способност да искористи велику количину контекста, па додатни сегменти изнад одређеног прага доприносе све мање. Поред тога, модел често може дати тачан одговор и на основу непотпуног контекста, ако су кључне информације присутне у бар неким од враћених сегмената. Коначно, метрике претраге мере само да ли су релевантни сегменти пронађени и како су рангирани, али не и колико је њихов садржај заиста користан за формирање одговора. Због свега тога, квалитет RAG система не може се поуздано проценити само на основу метрика претраге, већ захтева и непосредну оцену генерисаних одговора.

5.3 Ограничења експерименталне поставке

Резултате треба тумачити уз неколико ограничења експерименталне поставке.

Корпус се састоји искључиво од историјских чланака са Википедије на енглеском језику, па се закључци не могу без резерве пренети на друге домене, језике или типове докумената. Домен историје носи и своје специфичности, попут претежно чињеничног садржаја и наглашених хронолошких односа, какве други домени не морају делити.

Евалуациони скуп од 163 питања формиран је аутоматски, помоћу модела GPT-4o. Тиме је омогућено брзо добијање три типа питања, али тако генерисана питања могу одступати од оних која би поставили стварни корисници. Уз то, скуп ове величине ограничава статистичку снагу поређења.

Поређење стратегија сегментације додатно отежава мапирање релевантних сегмената, које уноси апроксимацију и тиме обара метрике претраге код стратегија различитих од базне — пре свега код малих сегмената и контекстуалне сегментације.

Коначно, кроз све експерименте задржани су исти генеративни модел и исти промпт. Другачији модел или другачије формулисан промпт могли би изменити резултате, нарочито у погледу тога колико успешно модел користи пронађени контекст.

5.4 Ограничења LLM као судија приступа

Поређење аутоматске и мануелне евалуације, приказано у одељку 4.8, показало је добру општу саглашеност, али и неколико систематских тенденција које ваља имати у виду. Аутоматска евалуација показује благу склоност ка строжем оцењивању потпуности и блажем оцењивању утемељености, што значи да њене оцене нису сасвим неутралне у односу на људске, већ носе препознатљив образац.

Уочено је и да аутоматска евалуација ређе препознаје халуцинације него човек. У применама где је откривање нетачних тврдњи важно, ово ограничење треба узети у обзир, па се аутоматска оцена у таквим случајевима не би смела користити као једини ослонац.

Коначно, иако је за евалуацију изабран модел из друге породице модела

(Claude Opus компаније Anthropic), чиме се умањује ризик од пристрасности према одговорима које генерише модел компаније OpenAI, тиме се не уклања могућност да и сам модел за евалуацију има одређене систематске склоности. Због свега наведеног, аутоматску евалуацију треба схватити као користан и скалабилан, али не и потпуно поуздан вид процене, који најбоље функционише уз повремену проверу људском евалуацијом.

Глава 6

Закључак

Спроведена анализа потврђује да квалитет RAG система не зависи од једне компоненте, већ од заједничког деловања сегментације докумената, репрезентације текста, броја враћених сегмената и начина претраге. Најбоље укупне резултате постигло је комбиновање претраге на основу векторских репрезентација и лексичке претраге са ренгирањем, док избор између хибридне и претраге на основу векторских репрезентација у пракси зависи од тога колика се важност придаје брзини одговора.

Резултати уједно откривају и компромисе који прате ове системе. Повећање броја враћених сегмената доноси највеће побољшање метрика претраге, али може благо нарушити утемељеност одговора, јер у контекст доспевају и сегменти који нису важни за питање. Насупрот томе, контекстуална сегментација постиже највишу утемељеност упркос слабијој претрази. Из оба запажања следи кључни налаз рада: боље метрике претраге не воде нужно и сразмерно бољим одговорима, па се квалитет RAG система не може поуздано проценити без непосредне оцене генерисаних одговора.

Поређење са мануелним оцењивањем показало је да аутоматска евалуација заснована на приступу LLM као судија даје оцене блиске људским. Ипак, она показује и препознатљиве склоности — потпуност оцењује строже, утемељеност блаже, а халуцинације уочава ређе у односу на мануелну евалуацију — па је треба схватити као скалабилну допуну људској процени, а не као њену замену.

Главни доприноси рада су систематично поређење конфигурација RAG система у контролисаним условима, анализа односа између квалитета претраге и квалитета генерисаних одговора, и валидација приступа LLM као судија

поређењем са мануелном евалуацијом.

Даљи рад може бити усмерен у неколико праваца: ка напреднијим стратегијама сегментације, укључујући семантичку и хијерархијску; ка примени више генеративних модела, ради провере у којој мери закључци зависе од избора модела; ка проширењу корпуса на друге домене и језике; и ка развоју метода за аутоматско подешавање параметара система према карактеристикама корпуса и типу питања.

Библиографија

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[2] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint*, 2024. arXiv:2402.03216.

[3] Chroma. Chroma: The open-source embedding database. `https://www.trychroma.com/`, 2024.

[4] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.

[5] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–163, 2024.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint*, 2024. arXiv:2312.10997.

[7] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint*, 2023. arXiv:2311.05232.

[8] Greg Kamradt. 5 levels of text splitting. `https://github.com/FullStackRetrieval-com/RetrievalTutorials`, 2023.

[9] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[11] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint*, 2022. arXiv:2201.10005.

[12] OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. arXiv:2303.08774.

[13] OpenAI. New embedding models and API updates. `https://openai.com/index/new-embedding-models-and-api-updates/`, 2024.

[14] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[16] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

Биографија аутора

Вук Стефановић рођен је 6. новембра 2000. године у Пожаревцу. Основну школу и гимназију завршио је у Великом Градишту, где је стекао Вукову диплому. Основне академске студије завршио је на Математичком факултету Универзитета у Београду, на студијском програму Информатика, где тренутно похађа мастер академске студије.

Током студија развио је интересовање за машинско учење, обраду природног језика, системе за проналажење информација и велике језичке моделе. Практично искуство стекао је радом на пројектима из области науке о подацима и вештачке интелигенције.