

Компаративна анализа одабраних приступа оптимизацији претраге у архитектури RAG

Comparative Analysis of Selected Retrieval Optimization Approaches in RAG Architecture

Сања Петровић, Факултет техничких наука, Нови Сад

Студијски програм – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – Унапред обучени језички модели често немају знање из уско специјализованих домена као што је фармација, а додатан изазов представља и употреба српског језика. Због тога је потребна архитектура генерисања потпомогнутог претрагом (енгл. *Retrieval-Augmented Generation (RAG)*), која генеративни велики језички модел проширује са базом екстерног знања. Постоје бројни приступи којима је могуће побољшати претрагу унутар архитектуре RAG, али је неопходно анализирати и оценити њихов утицај у контексту домена фармације на српском језику. Из тог разлога спроведен је експеримент компаративне анализе утицаја приступа филтрирања на основу метаподатака, хибридне претраге, диверзификације и трансформације упита, чији су резултати представљени у овом раду.

Кључне речи: Архитектура RAG, компаративна анализа, векторска база података, велики језички модел, модел уграђивања

Abstract – Pre-trained language models often lack knowledge of narrowly specialized domains such as pharmacy, and the use of the Serbian language poses an additional challenge. For this reason, a Retrieval-Augmented Generation (RAG) architecture is needed, which extends a generative large language model with an external knowledge base. There are numerous approaches for improving retrieval within a RAG architecture, but it is necessary to analyze and evaluate their impact in the context of the pharmacy domain in the Serbian language. Therefore, a comparative analysis experiment was conducted on the impact of metadata-based filtering, hybrid search, diversification, and query transformation approaches, the results of which are presented in this paper.

Keywords: Architecture RAG, comparative analysis, vector database, large language model, embedding model

НАПОМЕНА: Овај рад проистекао је из мастер рада чији ментор је био др Марко Вјештица, доцент.

1. УВОД

Последњих година забележен је раст развоја и примене вештачке интелигенције, при чему су велики језички модели (енгл. *Large Language Models (LLM)*) нашли примену у многим сферама савременог друштва и индустрије. Како су обучени на јавно доступним подацима, често не располажу тачним и ажурним информацијама из уско специјализованих домена. Једно решење за тај проблем је архитектура генерисања потпомогнуто претрагом (енгл. *Retrieval-Augmented Generation (RAG)*), која комбинује генеративни модел са базом екстерног знања у сврху проширења контекста модела.

Циљ овог мастер рада је да пружи опште смернице приликом одабира приступа оптимизацији претраге у домену фармације на српском језику. Садржај рада обухвата теоријске концепте архитектуре RAG, као и експеримент. У оквиру експеримента упоређено је више приступа оптимизацији претраге унутар система који имплементира архитектуру RAG при чему је систем пројектован као виртуелни асистент у улози фармацеутског стручњака који одговара на питања о лековима на српском језику.

2. ТЕОРИЈСКЕ ОСНОВЕ АРХИТЕКТУРЕ RAG

Архитектура RAG обухвата приступ са три међусобно повезане фазе: индексирање базе података, претрагу и генерисање одговора.

2.1. Фазе приступа архитектуре RAG

Индексирање базе података обухвата прикупљање и пречишћавање изворних, најчешће текстуалних података, њихову поделу на мање сегменте поступком сегментације, као и трансформисање сегмената у густе или ретке векторе који се чувају у векторској бази података [1].

Претрага базе података представља централну фазу архитектуре RAG, у којој претраживач кориснички упит трансформише у векторску репрезентацију истим моделом уграђивања примењеним при индексирању, а затим врши претрагу базе података ради проналажења

најрелевантнијих сегмената који се прослеђују генератору као контекст [1].

Генерисање представља завршну фазу у архитектури *RAG*, при чему се одговор генерише на основу допремљених сегмената који служе као контекст и системске инструкције понашања генератора [2].

2.2. Приступни оптимизације претраге у архитектури *RAG*

Трансформација упита је приступ оптимизацији претраге који подразумева модификацију корисничког упита пре прослеђивања претраживачу, а обухвата технике попут преформулисања упита, преформулисања на више упита, проширења упита, декомпозиције упита, уграђивања хипотетичких докумената и враћања упита корак уназад [2].

Филтрирање на основу метаподатака је приступ оптимизацији претраге која сужава простор претраге на основу структурираних атрибута придружених сегментима приликом индексирања [3].

Хибридна претрага комбинује ретке векторе, попут оних насталих функцијом најбољег подударања (енгл. *Best Matching 25 (BM25)*), и густе векторе тако што иницијално претражује базу података помоћу оба механизма, спаја резултате у јединствену листу, а затим их поново рангира [2].

Приликом претраживања комплексних докумената може се десити да неколико најбоље ранжираних сегмената садржи готово идентичне информације. Скуп резултата може се диверзификовати алгоритмом максималне маргиналне релевантности (енгл. *Maximal Marginal Relevance (MMR)*) [4].

2.3. Евалуација архитектуре *RAG*

Радни оквир за процену архитектуре *RAG* (енгл. *Retrieval-Augmented Generation Assessment (RAGAs)*) представљен је 2024. године [5]. Овај радни оквир дефинише више метрика, од којих су у овом раду представљене прецизност контекста, одзив контекста, верност и релевантност одговора.

Прецизност контекста процењује способност претраживача да релевантне сегменте рангира испред нерелевантних у листи преузетих резултата. Одзив контекста рачуна у којој мери је претраживач успео да преузме све информације неопходне за тачан одговор. Верност мери колико је генерисани одговор заснован на преузетом контексту. Релевантност одговора процењује у којој мери генерисани одговор директно и потпуно адресира постављени упит [6].

3. ПРИПРЕМА ЕКСПЕРИМЕНТА

Агенција за лекове и медицинска средства Србије (*АЛИМС*) је државна институција задужена за регулацију и документацију лекова и медицинских средстава унутар Републике Србије. У њиховој надлежности је, између осталог, да за сваки лек саставе Сажетак карактеристика лека који описује својства и

званично одobreне услове за употребу лека или медицинског средства.

Експеримент је пројектован на начин да одражава систем коме могу постављати питања стручна лица и пацијенти. Природа домена захтева високу тачност и поузданост одговора, али дозвољава чекање на одговор.

Унутар експеримента прате се представљене метрике прецизности и одзива контекста, верности и релевантности одговора како би се оценио квалитет претраге и генерисања одговора зависно од примењеног приступа оптимизацији претраге. Поред тога, експеримент обухвата и мерење финансијских трошкова и укупног времена потребно за извршавање сваког приступа оптимизацији претраге. Евалуација је спроведена над вештачки направљеним златним скупом података намењеним за тестирање.

3.1. Радно окружење

Уређај на којем је извршен експеримент је рачунар *MacBook M4 Pro*. Процесор се састоји од дванаест језгара. Капацитет радне меморије износи 24 GB, а максимална искоришћеност је ограничена на 12 GB. Рачунар поседује полупроводнички диск који је капацитета 512 GB.

3.2. Обрада изворног скупа података

Као изворни скуп података одабрани су Сажетци карактеристика лека. Обрада изворног скупа података подразумева три фазе: фазу преузимања, фазу трансформације и фазу сегментације.

Фаза преузимања података подразумева аутоматизовано прикупљање докумената и метаподатака са веб странице институције *АЛИМС* [7], чиме је преузето укупно 3779 докумената величине 868 MB, док се у фази трансформације ови подаци претварају у формат *Markdown*. Фаза сегментације користи приступ родитељ-дете, у складу са хијерархијском структуром Сажетка карактеристика лека, где је дете-сегмент најмања јединица текста, појединачни пасус или табела, а његов родитељ наслов или поднаслов унутар кога се налази.

3.3. Векторско уграђивање

За векторско уграђивање сегмената одабран је модел *voyage-context-3*. Осим густих вектора, у овом кораку креирају се и ретки вектори применом алгоритма *BM25*, а сви вектори се заједно са оригиналним садржајем и метаподацима сегмента чувају као равни индекси унутар једне колекције у векторској бази података *Qdrant*.

3.4. Претрага

У склопу експеримента, поред основне претраге, имплементирано је и неколико приступа оптимизацији претраге, а то су филтрирање на основу метаподатака, хибридна претрага са поновним рангирањем, диверзификација и трансформације упита које обухватају преформулисање, преформулисање на више упита, проширење, декомпозицију, уграђивање

хипотетичких докумената и враћање упита корак уназад. Ови приступи су одабрани након анализе домена и изворног скупа података, као и због међусобно сличног нивоа комплексности имплементације.

3.5. Генерисање одговора

Резултати претраге заједно са системском инструкцијом прослеђују се генератору који формулише коначан одговор на српском језику, при чему је одабран модел *gemi-3.1-flash-lite-preview*.

3.6. Евалуација

За потребе евалуације коришћен је радни оквир *RAGAs*. Процес евалуације се заснива на постојању златног скупа података који служи као референтни скуп парова питање-одговор. Имплементиран је сервис за евалуацију који процењује систем на основу датог питања, референтног одговора, пронађених сегмената и генерисаног одговора.

Одабране метрике за евалуацију су прецизност контекста, одзив контекста, верност и релевантност одговора. Евалуација захтева велики број позива апликативном интерфејсу модела судије, због чега је одабран велики језички модел *gemi-3.1-flash-lite-preview*.

Експеримент је подељен на два дела. Првенствено је тестирана контролна група *B* која чини основну варијанту имплементације без примене иједног приступа оптимизацији претраге, и варијанте које имплементирају приступе филтрирања на основу метаподатака (*F*), хибридне претраге са поновним рангирањем (*H*), диверзификације (*MMR*), као и њихове комбинације. Други део експеримента обухвата тестирање трансформације упита примењене над основном варијантом (*B*) и над најбољом варијантом из првог дела експеримента – комбинација филтрирања на основу метаподатака и хибридне претраге са поновним рангирањем, (*F + H*). Тестирања су вршена у следећем редоследу, обележено одговарајућим индексом: преформулисање (*r*), преформулисање на више упита (*m*), проширење (*e*), декомпозиција (*d*), уграђивање хипотетичких докумената (*h*) и враћање упита корак у назад (*s*). За сваки део експеримента се мере и трошкови на основу укупне цене претраге и генерисања одговора за сва питања из златног скупа података, као и укупног времена које је потребно да се изврше претрага и генерисања одговора.

Експеримент се понавља пет пута ради поузданости резултата. Резултати свих понављања се агрегирају и рачуна се средња вредност за сваку метрику по варијанти, заокружено на две децимале.

4. РЕЗУЛТАТИ ЕКСПЕРИМЕНТА

У наставку су дати резултати првог и другог дела експеримента у табеларном формату (Табеле 1 и 3), као и измерени трошкови и време извршавања упита (Табеле 2 и 4).

Табела 1. Резултати првог дела експеримента

Варијанта	Прецизност контекста	Одзив контекста	Верност	Релевантност одговора
<i>B</i>	0,81	0,60	0,83	0,77
<i>F</i>	0,81	0,71	0,89	0,85
<i>H</i>	0,91	0,67	0,85	0,84
<i>MMR</i>	0,97	0,54	0,78	0,80
<i>F + H</i>	0,87	0,77	0,94	0,92
<i>F + MMR</i>	0,92	0,59	0,86	0,91
<i>H + MMR</i>	0,93	0,59	0,85	0,89
<i>F + H + MMR</i>	0,93	0,59	0,85	0,89

Табела 2. Трошкови и време извршавања првог дела експеримента

Варијанта	Цена (\$)	Време извршавања (s)
<i>B</i>	0,04	601,69
<i>F</i>	0,05	266,07
<i>H</i>	0,10	582,68
<i>MMR</i>	0,07	635,54
<i>F + H</i>	0,09	427,84
<i>F + MMR</i>	0,07	319,46
<i>H + MMR</i>	0,08	673,91
<i>F + H + MMR</i>	0,07	334,66

Табела 3. Резултати другог дела експеримента

Варијанта	Прецизност контекста	Одзив контекста	Верност	Релевантност одговора
<i>B_r</i>	0,84	0,55	0,85	0,74
<i>B_m</i>	0,82	0,55	0,86	0,82
<i>B_e</i>	0,87	0,61	0,80	0,75
<i>B_d</i>	0,55	0,55	0,55	0,55
<i>B_h</i>	0,83	0,64	0,88	0,80
<i>B_s</i>	0,86	0,61	0,84	0,80
<i>(F + H)_r</i>	0,86	0,79	0,93	0,90
<i>(F + H)_m</i>	0,83	0,74	0,93	0,91
<i>(F + H)_e</i>	0,87	0,83	0,94	0,92
<i>(F + H)_d</i>	0,87	0,83	0,94	0,93
<i>(F + H)_h</i>	0,86	0,80	0,95	0,92
<i>(F + H)_s</i>	0,88	0,79	0,93	0,91

Табела 4. Трошкови и време извршавања другог дела експеримента

Варијанта	Цена (\$)	Време извршавања (s)
<i>B_r</i>	0,06	1003,94
<i>B_m</i>	0,07	1663,05
<i>B_e</i>	0,06	939,14
<i>B_d</i>	0,06	1596,73
<i>B_h</i>	0,06	1095,06
<i>B_s</i>	0,05	1047,04
<i>(F + H)_r</i>	0,14	583,20
<i>(F + H)_m</i>	0,26	699,05
<i>(F + H)_e</i>	0,16	695,68
<i>(F + H)_d</i>	0,22	723,57
<i>(F + H)_h</i>	0,17	747,30
<i>(F + H)_s</i>	0,14	712,98

5. ДИСКУСИЈА РЕЗУЛТАТА ЕКСПЕРИМЕНТА

Резултати експеримента показују да је за представљени систем филтрирање на основу метаподатака једна од значајнијих метода оптимизација претраге. У Табели 1 може се приметити да овај приступ оптимизацији претраге побољшава одзив контекста, верност и релевантност одговора, а Табела 2 показује да примена филтрирања на основу метаподатака скраћује и време извршавања. Природа домена система, као и изворног скупа докумената, се добро слаже са овим приступом, с обзиром да се може очекивати да значајан број постављених питања буде у вези са конкретним леком или активном супстанцом. Ограничење филтрирања на основу метаподатака је то

да се овај приступ не би могао применити на генеричка питања.

Хибридна претрага са поновним рангирањем такође побољшава квалитет претраге, посебно у комбинацији са филтрирањем на основу метаподатака, што се види у Табели 1. Разлог за побољшање успешности налази се у томе што систем ради са специјализованим терминима који се ретко налазе у вокабулару било ког модела опште намене, а ефекат овога је додатно изражен чињеницом да су подаци на српском језику.

Посматрајући резултате приказане у Табели 1, видљиво је да је приступ диверзификације у значајној мери допринео побољшању метрике прецизности контекста, али по цену погоршања метрике одзива контекста, што у осетљивом домену као што је фармација није погодно.

Када су у питању трансформације упита, резултати експеримента приказаних у Табели 3 показују да је утицај упита на претрагу ограничен уколико само претраживање има много шума. Пре трансформисања упита, препоручљиво је оптимизовати претрагу другим методама као што су филтрирање на основу метаподатака и хибридна претрага са поновним рангирањем. Такође, Табела 4 показује да трансформације упита повећавају трошкове и време извршавања, па би се препоручио опрез при одабиру овог приступа оптимизацији претраге.

6. ЗАКЉУЧАК

У овом раду представљено је поређење одабраних приступа оптимизацији претраге у архитектури *RAG* унутар домена фармације на српском језику.

Резултати експеримента наговештавају да архитектура *RAG* може бити погодан избор за претрагу великих скупова података унутар домена фармације на српском језику. Квалитет имплементације се може повећати уз примену одговарајућих приступа оптимизацији претраге као што су филтрирање на основу метаподатака и хибридна претрага густих и ретких вектора са поновним рангирањем, уз које у одређеним случајевима може бити корисно применити и трансформације упита.

Будући правац истраживања обухватио би проширење експеримента компонентом која ће на основу упита закључити да ли ће и на који начин извршити трансформацију упита. Додатно, потребно је интегрисати и графску базу података која би могла да побољша квалитет одговора на питања која обухватају разумевање комплексних веза између различитих ентитета. Такође, фино подешавање модела уграђивања би потенцијално омогућило боље резоновање специфичне терминологије из домена фармације на српском језику.

7. ЛИТЕРАТУРА

- [1] A. Bansal, "Optimizing RAG with Hybrid Search and Contextual Chunking," *Journal of Engineering and Applied Sciences Technology*, vol. 5, 2023, doi: 10.47363/JEAST/2023(5)E114.
- [2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.
- [3] M. Poliakov and N. Shvai, "Multi-Meta-RAG: Improving RAG for Multi-Hop Queries using Database Filtering with LLM-Extracted Metadata," in *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*, Cham: Springer Nature Switzerland, 2024, doi: 10.1007/978-3-031-81372-6_25.
- [4] J. G. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, 1998, doi: 10.1145/290941.291025.
- [5] S. Es, J. James, L. E. Anke, and S. 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*, 2024, doi: 10.18653/v1/2024.eacl-demo.16.
- [6] https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/ (приступљено у мају 2026.)
- [7] <https://www.alims.gov.rs/humanilekovi/pretrazivanje-humanih-lekova/> (приступљено у мају 2026.)

Кратка биографија:

Сања Петровић рођена је у Алексинцу 2001. године. Основне академске студије уписала је 2019. године на Факултету техничких наука, студијски програм Рачунарство и аутоматика, који је завршила 2023. године. Након тога, уписала је мастер академске студије на Факултету техничких наука, студијски програм Рачунарство и информатика, стручна област Рачунарство високих перформанси.

Професионалну каријеру започела је 2022. године као софтверски инжењер, а од 2023. године ради на позицији инжењера података.

Контакт: petroviccsanja@gmail.com