COMPARATIVE EVALUATION OF CHUNKING AND RETRIEVAL PARAMETERS IN A RAG-BASED SYSTEM

TEODORA T. STOYANOVA, STOYAN R. STOYANOV, ATILA B. AKIFOV, RADOSTIN E. RAFAILOV

ABSTRACT: *This paper investigates the influence of chunking and retrieval parameters on the performance of a RAG-based system under different user query formulations. The experiment is conducted using a university regulatory document, with two chunking configurations and two values for the number of retrieved chunks. A total of 30 queries are organized into 10 thematic groups, with each group containing three semantically equivalent queries formulated in direct, conversational, and paraphrased forms. Both the retrieved context and the generated response are evaluated by considering the position of relevant information, the coverage of predefined facts, and robustness to changes in query formulation.*

KEYWORDS: *Retrieval-Augmented Generation (RAG), text chunking, information retrieval, large language models*

DOI: https://doi.org/10.46687/MWRY6932

СРАВНИТЕЛНА ОЦЕНКА НА ПАРАМЕТРИ ЗА СЕГМЕНТИРАНЕ И ИЗВЛИЧАНЕ В RAG-БАЗИРАНА СИСТЕМА*

ТЕОДОРА Т. СТОЯНОВА, СТОЯН Р. СТОЯНОВ, АТИЛА Б. АКИФОВ, РАДОСТИН Е. РАФАИЛОВ

АБСТРАКТ: *В статията се изследва влиянието на параметрите за сегментиране и извличане върху работата на RAG-базирана система при различни формулировки на потребителските заявки. Експериментът е проведен върху нормативен документ от университетска среда с две конфигурации за сегментиране и две стойности за броя на извличаните фрагменти. Използвани са 30 заявки, организирани в 10 тематични групи, всяка от които включва директна, разговорна и парафразирана формулировка на едно и също информационно търсене. Оценяват се както извлеченият контекст, така и генерираният отговор, като се проследяват позицията на релевантната информация, покритието на предварително определени факти и устойчивостта при промяна на формулировката.*

1 Въведение

Големите езикови модели (Large Language Models, LLMs) през последните години повишиха ефективността си, но използването на знание, съхранено единствено в параметричната памет на модела, е свързано с редица ограничения. Сред тях са трудното актуализиране на усвоените знания, рискът от генериране на неверни твърдения и ограничената достъпност до специфична информация [1].

Генерирането с допълнително извличане на информация (Retrieval-Augmented Generation, RAG) преодолява част от тези ограничения чрез съчетаване на генеративните възможности на езиковите модели с информация, извличана от външни източници. Това позволява на системата да използва обновяема и специфична за конкретната област

* Настоящата статия е частично финансирана по проект № РД-08-88/02.02.2026 г.

информация [2]. Но използването на външни източници не гарантира надеждността на генерирания отговор.

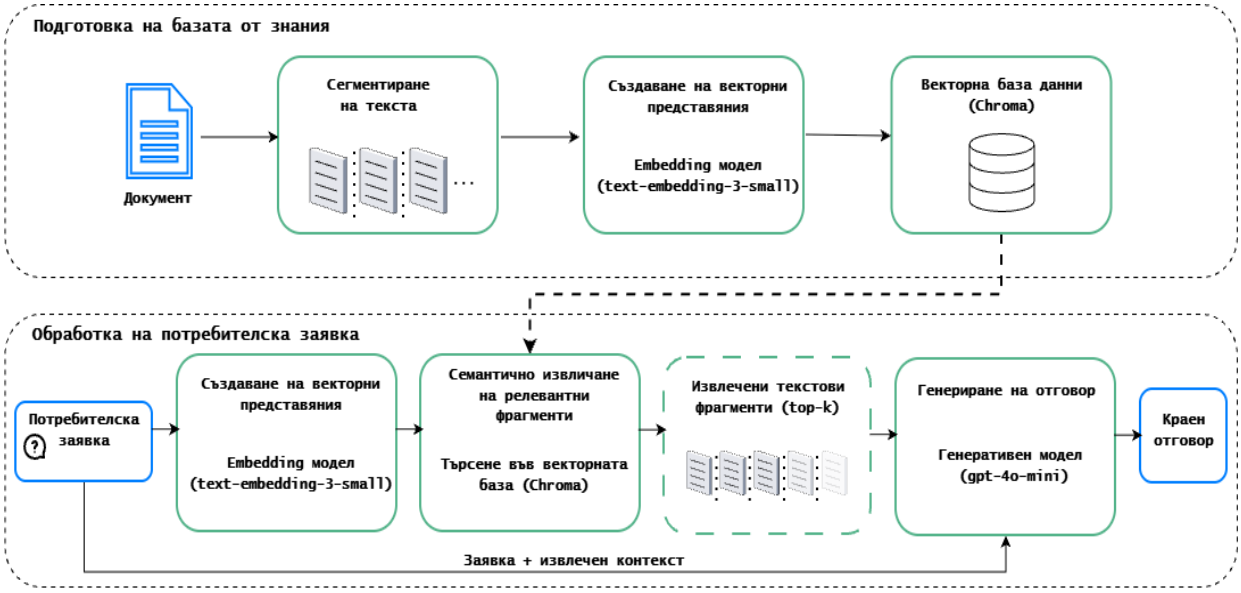
За да използва необходимата информация, системата трябва първо да я открие и извлече въз основа на запитването на потребителя. Въпросът може да бъде зададен директно, да се използва разговорна формулировка или да се парафразира, без да се променя неговото значение. Това може да доведе до различен извлечен контекст и съответно до различно качество на крайния отговор. При работа със структурирани и нормативни документи това е от особено значение.

Настоящата статия е насочена към анализ на устойчивостта на RAG система при семантично еквивалентни, но различно формулирани потребителски заявки. Основният изследователски въпрос е как сегментирането на документите и броят на извличаните фрагменти влияят върху способността на системата да запазва качеството на информационното извличане и на генерирания отговор. За целта се анализират както резултатите от извличане, така и крайните отговори на системата.

2 Архитектура и организация на експеримента

2.1. RAG архитектура и база знания

Процесите в реализираната RAG система са: подготовка и индексиране на базата знания и обработка на потребителската заявка (фиг. 1). В първия, избраният документ се разделя на фрагменти (chunks), които се преобразуват във векторен вид и се съхраняват във векторна база данни. При постъпване на потребителска заявка тя също се преобразува и чрез семантично търсене се извличат най-близките до нея фрагменти. Те се включват към заявката, след което езиковият модел генерира крайния отговор. Тази последователност съответства на основните етапи на RAG – индексиране, извличане и генериране [2].



Фигура 1. Последователност на обработката в използваната RAG система

За външна база знания е използван документът *Правилник за структурата и организацията на процеса на обучение в Шуменския университет „Епископ Константин Преславски“*. Информацията в него е организирана в йерархична структура от глави, членове, алинеи и точки. За използването му от RAG системата текстът е преработен в Markdown формат със запазена логическа структура.

За преобразуването на текстовите фрагменти във векторен вид е използван моделът *text-embedding-3-small*, а векторите са съхранени в Chroma база данни. Генерирането на крайния отговор се извършва чрез *gpt-4o-mini*. За коректното експериментално оценяване е изключена опцията за история на разговора и по този начин заявките се обработват независимо една от друга.

2.2. Сегментиране и извличане

Сегментирането определя начина, по който документът се разделя на отделни текстови единици, за които впоследствие се създават векторни представяния [3, 4]. За процеса на извличане се променят два параметъра: конфигурацията за сегментиране на базата знания и броят на извличаните фрагменти (top-k).

В изследването се използват следните две конфигурации за дължина на текстовия фрагмент и припокриването между тях, което е 20% от фрагмента: *chunk_size* = 500 с *chunk_overlap* = 100 и *chunk_size* = 1000 с *chunk_overlap* = 200. Стойностите са определени спрямо структурата на използвания нормативен документ.

Вторият променян параметър определя броя на най-високо класираните фрагменти, които се предоставят като контекст на генеративния модел. В изследването се използват две негови стойности – *top-k* = 3 и *top-k* = 5. Така при една и съща конфигурация за фрагментиране се използва един и същ индекс, като се променя единствено броят на фрагментите, включени в контекста на генеративния модел. По този начин се формират две конфигурации на сегментиране – 500/100 и 1000/200, всяка от които се изпълнява при двете стойности на *top-k*.

2.3. Експериментални конфигурации

Формираните четири експериментални конфигурации са представени в *Таблица 1*. Останалите основни компоненти на RAG системата се запазват непроменени, така че резултатите да могат да бъдат съпоставени при еднакви условия. Подходът, при който отделни параметри се изменят контролирано при фиксиране на останалите настройки, се използва и в експериментални изследвания на RAG системи [5].

Таблица 1. Експериментални конфигурации

Конфигурация	chunk_size	chunk_overlap	top-k
C1	500	100	3
C2	1000	200	3
C3	500	100	5
C4	1000	200	5

C1 и C3 използват индекса, създаден при *chunk_size* = 500 и *chunk_overlap* = 100, а C2 и C4 – индекса при *chunk_size* = 1000 и *chunk_overlap* = 200. Разликата между C1 и C3, както и между C2 и C4, е стойността на *top-k*. По този начин за четирите експериментални конфигурации са необходими две отделно индексирани колекции във векторната база данни. За всяко изпълнение се записват използваната конфигурация, извлечените текстови фрагменти, позицията им в класирането, стойността на разстоянието (distance) между векторното представяне на заявката и съответния фрагмент и генерираният отговор.

Експерименталният набор е организиран в 10 тематични групи. Всяка група съдържа три семантично еквивалентни формулировки – директна *(Q1)*, разговорна *(Q2)* и парафразирана *(Q3)*, което формира 30 заявки. Всяка от тях се изпълнява при четирите конфигурации от C1 до C4, т. е. общо 120 независими експериментални изпълнения.

3 Методика за оценяване на RAG системата

3.1. Набор от заявки и референтни данни

Трите варианта по всяка тема (Q1, Q2 и Q3) се оценяват спрямо един и същ референтен отговор. Този подход позволява да се изследва устойчивостта на системата при промяна на формулировката на дадена заявка, без да се променя търсената информация [6].

Включените в изследването тематични групи се различават по сложността на очаквания отговор. Част от въпросите изискват извличане на единичен конкретен факт, докато други предполагат съчетаване на няколко взаимосвързани факта от документа. Въпросите се различават по своята сложност, но влиянието на този фактор не е изследвано.

За всяка тема предварително е определена фактическа основа, извлечена от използвания нормативен документ. Тя включва референтен отговор, точна препратка към съответната разпоредба и набор от съществени факти (essential facts), които трябва да присъстват в пълния коректен отговор. Част от тях са определени като основни факти (core facts) – минималната информация, необходима за предаване на основния правилен смисъл [7]**.** В настоящото изследване тези факти са зададени предварително, а не се генерират автоматично от езиков модел.

3.2. Процедура и показатели за оценяване

Провеждането и оценяването на експериментите са автоматизирани с помощта на два програмни скрипта. В първия етап всички заявки се изпълняват при четирите експериментални конфигурации (C1 - C4). Във втория етап получените резултати се съпоставят с предварително зададените референтни факти и се изчисляват показателите за извличането и за крайния отговор. Разделното оценяване на извлечения контекст и генерирания отговор е в съответствие с общия подход за оценяване на RAG системи като RAGAS [8] и ARES [9].

На ниво извличане се отчита дали сред върнатите фрагменти присъства релевантна информация и на каква позиция е намерен първият релевантен фрагмент. Показателят *успешно извличане* (retrieval_success) приема стойност 1, когато в извлечения набор има поне един релевантен фрагмент, и 0, ако няма. *Позиция на първия релевантен фрагмент* (first_relevant_rank) показва мястото му в класирането. *Средният реципрочен ранг MRR* (Mean Reciprocal Rank) показва колко напред сред извлечените фрагменти се намира първият релевантен фрагмент, като по-високата стойност означава по-предна средна позиция.

Допълнително се изчислява *покритие на съществените факти при извличането* (retrieval_fact_coverage), което показва каква част от предварително зададените съществени факти присъстват в извлечения контекст, както и *покритие на основните факти при извличането* (retrieval_core_pass), което отчита дали всички основни факти са налични в него.

При крайния отговор автоматичната оценка проверява наличието на съществените факти чрез предварително зададени правила и допустими лексикални варианти. Всеки отговор се класифицира като *правилен* (correct), *частично правилен* (partial) или *неправилен* (incorrect). Показателят *покритие на съществените факти в отговора* (answer_fact_coverage) измерва дела на установените съществени факти, а следващият - *покритие на основните факти в отговора* (answer_core_pass) показва дали крайният отговор съдържа цялото фактическо ядро.

Чрез тези показатели се оценява както дали необходимите факти присъстват в извлечения контекст и на каква позиция се намира първият релевантен фрагмент, така и доколко съществените и основните факти са отразени в генерирания отговор. Получените за всяко изпълнение стойности се обобщават автоматично по експериментални конфигурации.

3.3. Оценяване на устойчивостта

Системата се приема за устойчива по даден критерий, когато качеството се запазва и при трите семантично еквивалентни формулировки. Тази постановка следва идеята за устойчивост към вариации на заявката (query variation robustness) [6].

Използват се три критерия за оценяване: устойчивост на пълния отговор (трите отговора са класифицирани като правилни), устойчивост на фактическото ядро (трите отговора съдържат всички основни факти) и устойчивост на извличането (необходимите основни факти присъстват в извлечения контекст и при трите формулировки). Резултатите се обобщават и според вида на заявката (директна, разговорна и парафразирана), с цел да се сравни поведението на системата при различни начини на изразяване на едно и също информационно търсене.

Оценяването се извършва автоматично въз основа на предварително зададени правила за разпознаване на съществените и основните факти, без да се използва допълнителен езиков модел за оценка на отговорите. За всеки предварително определен факт са зададени един или повече текстови шаблони, чрез които той се разпознава в извлечения контекст и в генерирания отговор. Така едни и същи критерии се прилагат към всички 120 експериментални заявки. Тъй като предварително зададените правила не могат да обхванат всички възможни формулировки, отговорите, определени автоматично като частично правилни или неправилни, се проверяват допълнително ръчно.

4 Резултати и анализ

4.1. Влияние на сегментирането и броя на извличаните фрагменти

Таблица 2 обобщава резултатите за четирите експериментални конфигурации чрез показатели, определящи два последователни етапа на RAG системата – извличането на информация и генерирането на крайния отговор. Показателите *правилни, частични, неправилни, строга точност и покритие на фактическото ядро* се отнасят до генерираните отговори. *Успешно извличане, MRR и Покритие на съществените факти при извличане* оценяват извлечения контекст преди генерирането на отговора. По този начин таблицата осигурява възможност за проследяване, както на успешното извличане на необходимата информация, така и на степента, в която тя е интегрирана и представена в крайния отговор.

Строгата точност показва дела на напълно правилните отговори, а покритието на *основните факти* – дела на отговорите, които съдържат цялото необходимо фактическо ядро, дори когато липсват някои допълнителни факти. *Успешното извличане* показва в какъв дял от случаите сред извлечените фрагменти присъства поне един релевантен резултат. *Средният реципрочен ранг* (MRR) характеризира позицията на първия релевантен фрагмент, като по-високата стойност означава по-предно класиране. *Фактическото покритие при извличането* показва каква част от предварително зададените съществени факти присъстват в извлечения контекст.

Основните резултати за четирите конфигурации са представени в Таблица 2. Строгата точност е между 33,33% и 40,00%, докато покритието на основните факти в крайния отговор е значително по-високо – между 83,33% и 86,67%. Това показва, че

голяма част от отговорите съдържат фактическото ядро, дори когато не покриват всички предварително зададени съществени факти.

Таблица 2: Основни резултати

Конф.	Правилни	Частични	Неправилни	Строга точност	Основни факти	Успешно извличане	MRR	Факт. покритие при извличане
C1	12	16	2	40,00%	86,67%	93,33%	0,7944	90,42%
C2	10	16	4	33,33%	86,67%	90,00%	0,7444	90,69%
C3	11	16	3	36,67%	83,33%	96,67%	0,8011	94,17%
C4	10	17	3	33,33%	86,67%	93,33%	0,7511	94,44%

При еднакъв брой извличани фрагменти конфигурацията 500/100 поставя първия релевантен резултат по-високо от 1000/200: средният реципрочен ранг (MRR) е 0,7944 спрямо 0,7444 при *top-k* = 3 и 0,8011 спрямо 0,7511 при *top-k* = 5. Разликите във фактическото покритие на извлечения контекст обаче са малки. Поради това резултатите не показват общо превъзходство на единия размер на фрагментите, а различие основно в позицията на релевантния фрагмент спрямо останалите.

Увеличаването на броя на извличаните фрагменти от 3 на 5 повишава успешното извличане и фактическото покритие и при двете конфигурации за сегментиране. Подобрението при извличането не води до подобрение на крайните отговори. Строгата точност не показва устойчиво нарастване, а броят на неправилните отговори не следва определена тенденция. Следователно в рамките на експеримента предоставянето на повече релевантна информация не гарантира по-пълен краен отговор.

4.2. Влияние на формулировката на заявката

Най-устойчиви са директните формулировки (Q1). При тях покритието на основните факти е 100% и няма неправилни отговори при нито една от четирите конфигурации. При разговорните (Q2) и парафразираните (Q3) заявки резултатите са по-променливи, като при Q3 делът на неправилните отговори достига 30% при C2 и C4 (Таблица 3).

Таблица 3: Резултати според вида на формулировката на заявката

Конфигурация	Формулировка	Строга точност	Основни факти	Неправилни
C1	Директна (Q1)	50,0%	100,0%	0,0%
C1	Разговорна (Q2)	30,0%	80,0%	10,0%
C1	Парафразирана (Q3)	40,0%	80,0%	10,0%
C2	Директна (Q1)	40,0%	100,0%	0,0%
C2	Разговорна (Q2)	40,0%	90,0%	10,0%
C2	Парафразирана (Q3)	20,0%	70,0%	30,0%
C3	Директна (Q1)	50,0%	100,0%	0,0%
C3	Разговорна (Q2)	30,0%	70,0%	20,0%
C3	Парафразирана (Q3)	30,0%	80,0%	10,0%
C4	Директна (Q1)	40,0%	100,0%	0,0%
C4	Разговорна (Q2)	40,0%	90,0%	0,0%
C4	Парафразирана (Q3)	20,0%	70,0%	30,0%

Резултатите показват, че семантично еквивалентните формулировки не водят непременно до еднакво поведение на системата. Наблюдаваната чувствителност към начина на формулиране е в съответствие с резултати от изследвания върху устойчивостта към вариации на заявките [2, 6]. Поради ограничения брой теми резултатите не позволяват извода, че всеки парафразиран въпрос е по-труден от разговорния, а е заключение отнасящо се до настоящото изследване.

4.3. Анализ на конкретен случай при различни формулировки на заявката

Наблюдаваното влияние на формулировката може да се проследи конкретно при тема 2 - *Условия за допускане до изпит* и конфигурация C3, характеризираща се с размер на фрагмента 500, припокриване 100 и брой на извлечените фрагменти 5 (Таблица 4). Трите заявки изразяват едно и също информационно търсене, но са формулирани съответно в директен, разговорен и парафразиран стил. Необходимата информация се съдържа във фрагмента, според който студентът има право да се яви на изпит по съответната дисциплина, ако е изпълнил заложените в учебната програма условия за допускане, посочени в информационния „Комплект на студента“ и публикувани в сайта на университета.

Таблица 4. Резултати при различните формулировки на заявката за Тема 2 и конфигурация C3

Вариант	Заявка	Резултат от извличането	Фактическо ядро в извлечения контекст	Краен отговор	Оценка
директна	„При какви условия студентът има право да се яви на изпит по дадена дисциплина?“	Правилният фрагмент за условията за допускане до изпит е на ранг 1	Да	Посочва, че трябва да са изпълнени условията от учебната програма, описани в информационния „Комплект на студента“	Правилен
разговорна	„Какво трябва да съм изпълнил, за да ме допуснат до изпит?“	На рангове 1 и 2 са фрагменти, свързани с държавни изпити; правилният фрагмент е на ранг 3	Да	Посочва изпълнение на задълженията по учебния план и полагане на всички семестриални изпити, т.е. условия за допускане до държавна изпитна сесия	Неправилен
парафразирана	„Какви предварителни изисквания трябва да бъдат изпълнени, за да може студентът да участва в семестриалното оценяване по съответната дисциплина?“	На ранг 1 е фрагмент за адаптирани условия при изпитване, на ранг 2 - за повишаване на оценка, а правилният фрагмент е на ранг 3	Да	„Информацията не е налична в предоставения документ.“	Неправилен

Примерът показва, че промяната във формулировката на едно и също информационно търсене води до различно подреждане на извлечените фрагменти. При директната заявка Q1 необходимият фрагмент е класиран на първа позиция и системата генерира правилен отговор. При разговорната и парафразираната формулировка същият фрагмент се измества на трета позиция, като пред него се класират други текстове, тематично свързани с изпити и оценяване, но нерелевантни спрямо конкретното информационно търсене.

Особено показателно е, че и при трите формулировки фактическото ядро присъства в извлечения контекст, но правилен краен отговор е получен само при Q1. При Q2 моделът използва информацията от по-високо класиран фрагмент и формулира отговор за допускане до държавна изпитна сесия, а при Q3 не използва наличния правилен фрагмент и отговаря, че информацията не е налична в документа. Следователно наличието на необходимата информация сред извлечените фрагменти само по себе си не гарантира правилен краен отговор.

Този случай илюстрира и необходимостта извличането и генерираният отговор да бъдат оценявани отделно. Наблюдаваната неустойчивост не се изразява единствено в това дали релевантната информация е извлечена, а и в позицията ѝ сред предоставените на генеративния модел фрагменти и начина, по който моделът използва получения контекст. Така примерът конкретизира установената при обобщените резултати разлика между устойчивостта на извличането и устойчивостта на крайния отговор.

4.4. Устойчивост и връзка между извличането и крайния отговор

Устойчивостта се оценява на ниво тематична група, като съответният критерий трябва да бъде изпълнен и при трите формулировки на заявката – директна, разговорна и парафразирана. Разглеждат се три критерия: устойчивост на пълния отговор, устойчивост на фактическото ядро и устойчивост на извличането. Резултатите за четирите експериментални конфигурации са представени в Таблица 5.

Таблица 5: Устойчивост на системата

Конфигурация	Строга устойчивост	Устойчивост на фактическото ядро	Устойчивост на извличането
C1	3/10	8/10	8/10
C2	1/10	7/10	8/10
C3	2/10	7/10	8/10
C4	1/10	6/10	8/10

Най-ниски са стойностите за устойчивостта на пълния отговор. Само при C1 три от десетте тематични групи съдържат правилни отговори и за трите формулировки, докато при C2 и C4 това условие е изпълнено само за една тема, а при C3 – за две.

По-високи са резултатите за устойчивостта на фактическото ядро. Тя е постигната при 8 от 10 теми за C1, при 7 от 10 за C2 и C3 и при 6 от 10 за C4. Това означава, че дори когато отговорите не съдържат всички предварително определени съществени факти, основните факти по-често се запазват и при трите формулировки на заявката.

Най-еднородни са резултатите при устойчивостта на извличането. За всяка от четирите конфигурации при 8 от 10 тематични групи основните факти присъстват в извлечения контекст и при трите формулировки. Увеличаването на броя на извличаните фрагменти от 3 на 5 не води до промяна на този резултат, тъй като стойността остава 8/10 както при C1 и C2, така и при C3 и C4.

Съпоставянето на трите критерия показва, че устойчивостта е по-висока на ниво извличане и фактическо ядро, отколкото при изискването за пълнота на крайния отговор. Следователно различната формулировка на едно и също информационно търсене оказва по-силно влияние върху пълнотата на генерирания отговор, докато извличането на основните факти остава по-устойчиво в рамките на изследваните конфигурации.

5 Заключение

Резултатите от проведеното изследване показват различия в поведението на RAG системата в зависимост от използваната конфигурация и начина на формулиране на

заявката. При еднакъв брой извличани фрагменти конфигурацията с по-малък размер на фрагмента осигурява по-предна позиция на първия релевантен резултат. Увеличаването на броя на извличаните фрагменти подобрява успешното извличане и фактическото покритие, но не води до съответно подобрение на крайните отговори. Различия се наблюдават и при трите вида формулировки, като директните заявки показват по-добри резултати от разговорните и парафразираните. По-високата устойчивост на извличането спрямо устойчивостта на пълния отговор показва, че наличието на необходимата информация в извлечения контекст не е достатъчно условие за получаването на пълен и правилен краен отговор.

Получените резултати подчертават необходимостта при оценяването на RAG системи да се разглеждат както качеството на извличането, така и крайният генериран отговор. Това е особено важно при използването на системи с изкуствен интелект в образователна среда, където прозрачността и разбираемостта на резултатите имат значение за доверието и ефективното им приложение [10]. Изследването е ограничено до един нормативен документ и сравнително малък набор от тематични групи, поради което бъдещата работа може да включи различни по структура документи, по-голям брой заявки и допълнителни стратегии за сегментиране и извличане.

ЛИТЕРАТУРА:

[1] Lewis, P., Perez, E., Piktus, A. et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

[2] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H., Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997v5, 2024.

[3] Günther, M., Mohr, I., Williams, D.J., Wang, B., Xiao, H., Late Chunking: Contextual Chunk Embeddings Using Long-Context Embedding Models. arXiv:2409.04701v3, 2025.

[4] Qu, R., Tu, R., Bao, F., Is Semantic Chunking Worth the Computational Cost? Findings of the Association for Computational Linguistics: NAACL 2025, pp. 2155–2177, 2025.

[5] Knollmeyer, S., Caymazer, O., Koval, L., Akmal, M. U., Asif, S., Mathias, S. G., & Großmann, D. (2026). Evaluating Retrieval Augmented Generation: A Comprehensive Review of Evaluation Dimensions, Question Types, and Application. SN Computer Science, 7, 576. https://doi.org/10.1007/s42979-026-05134-x

[6] Perçin, S., Su, X., Syed, Q.S., Howard, P., Kuvshinov, A., Schwinn, L., Scholl, K.-U., Investigating the Robustness of Retrieval-Augmented Generation at the Query Level. Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pp. 439–457, 2025.

[7] Pradeep, R., Thakur, N., Upadhyay, S., Campos, D., Craswell, N., Soboroff, I., Dang, H.T., Lin, J., The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models. Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), Padua, Italy, 2025. https://doi.org/10.1145/3726302.3730090

[8] Es, S., James, J., Espinosa-Anke, L., Schockaert, S., RAGAS: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp. 150–158, 2024.

[9] Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M., ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 338–354, 2024.

[10] Ivanova, D. N. (2025). Explaining Artificial Intelligence in Education: Approaches to Integration in Automated Information Systems. Journal Scientific and Applied Research, Vol. 29, pp. 194–203. ISSN 1314-6289 (Print), ISSN 2815-4622 (Online).

Теодора Стоянова
Шуменски университет „Епископ Константин Преславски“
t.stoyanova@shu.bg

Стоян Стоянов
Шуменски университет „Епископ Константин Преславски“
s.stoyanov@shu.bg

Атила Акифов
Студент, специалност „Софтуерно инженерство“
2632550009@shu.bg

Радостин Рафаилов
Шуменски университет „Епископ Константин Преславски“
r.rafailov@shu.bg