БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Мирослава Дончева Димитрова

РАМКА ЗА ОЦЕНКА НА СЪДЪРЖАНИЕ, ГЕНЕРИРАНО С РАЗШИРЕНО ИЗВЛИЧАНЕ

АВТОРЕФЕРАТ

на дисертация
за придобиване на образователната и научна степен „доктор“

по докторска програма „Информатика“
Професионално направление: 4.6. „Информатика и комютърни науки“

Научен ръководител:
Академик Иван Попчев

София, 2026 г.

Дисертационният труд е обсъден и допуснат до защита на разширено заседание на секция “_____”
на ИИКТ–БАН, проведена на _____ 2026 г.

Дисертационният труд е структуриран в увод, 5 глави, заключение, апендикс, списък на публикациите, списък на забелязаните цитирания, декларация за оригиналност на резултатите и библиография. Дисертационният труд е в общ обем от 215 страници, 24 фигури, 45 таблици, 34 формули, и 145 литературни източника.

Защитата на дисертацията ще се състои на _____ 2026 в _____ часа в зала _____, на блок 2 на ИИКТ–БАН на открито заседание на научно жури, в състав:

Научно жури:

1. _____
2. _____
3. _____
4. _____
5. _____

Рецензиите и становищата на членовете на научното жури и авторефератът са публикувани на сайта на ИИКТ–БАН.

Материалите за защитата са на разположение на интересуващите се в стая _____ at ИИКТ–БАН, ул. Акад. Г. Бончев, блок 2.

Автор:

Мирослава Димитрова

Заглавие:

РАМКА ЗА ОЦЕНКА НА СЪДЪРЖАНИЕ,
ГЕНЕРИРАНО С РАЗШИРЕНО ИЗВЛИЧАНЕ

УВОД

Езиковите модели на основата на трансформери [1], при които броят на параметрите достига милиарди, по-познати като големи езикови модели (Large Language Models, LLMs), намират широко приложение в обработката на естествен език (Natural Language Processing, NLP) при задачи като машинен превод, обобщаване, отговаряне на въпроси и водене на диалог [2]. Въпреки че генерират граматично коректен и свързан текст, надеждността им остава ограничена в ситуации, в които са необходими конкретни факти, актуална информация или проследими източници [3]. Когато отговорите трябва да бъдат подкрепени с външни доказателства, а не да се извеждат единствено от параметричното знание на модела, тези модели често генерират необосновани или остарели твърдения [4]. Подходът Retrieval-Augmented Generation (RAG) преодолява това ограничение, като отделя извличането на информация от генерирането на текст [5].

Вместо да разчита изцяло на параметрично знание, RAG извлича информация от външен корпус и генерира отговор въз основа на извлечения контекст. Тази архитектура позволява знанието да се актуализира чрез обновяване на корпуса, без моделът да се обучава повторно, и подпомага генерирането на отговори, основани на източници, когато такива са налични. Ефективността на обвързването с източници обаче зависи от политиката на извличане: как се определя съответствието, какъв контекст се подбира и как прагът на сходство и начинът на подреждане на резултатите оформят информацията, предоставена на генератора [6]. Тя зависи също от способността на генератора да използва извлечения контекст последователно, вместо да го "пренаписва" с неподвърдено съдържание. Затова при оценяването на подобни системи е необходимо да се разглеждат не само плавността и свързаността на текста, но и влиянието на извличането и изборът на контекст върху фактическата достоверност и пълнотата на отговора.

Актуалност на темата

RAG системите намират все по-широко приложение в области, където фактическата точност и проследимостта пряко влияят върху качеството на вземане на решения, включително в здравеопазването [7] и правните изследвания [8]. В такива условия, неверни твърдения носят практически последици: неправилно клинично указание или измислен правен прецедент, могат да компрометират последващи решения. Пропуските при извличането, когато релевантни доказателства съществуват, но не са извлечени - могат да обезсмислят ползите от използването на външен контекст.

Въпреки че някои големи езикови модели, разполагат с функции за уеб търсене, подобна функционалност не осигурява по своята същност достоверност на източниците при контролирани условия на оценка. Уеб ресурсите се характеризират с непостоянна стабилност, достъпът до тях невинаги е налице, а критериите за подбор или съхранение на съдържанието често не са прозрачни — фактори, които затрудняват сравнението между отделните експериментални цикли. Чисто параметричните модели, също така, остават склонни да генерират убедителни, но фактологично грешни отговори, когато източниците липсват, противоречат си илиса слабо свързани със заявката. Освен това, те често предоставят ограничена или неясна информация за източника на отделните твърдения [3], [9].

Мотивация на изследването

Въпреки че обзорните публикации отчитат бързо развитие на архитектурите и работните процеси в RAG системите, те също подчертават липса на единен подход в конфигурациите и практиките за оценка. Това ограничава възможностите за съпоставимост и вземането на информирани решения при избор на модел за внедряване [10]. Изследването е мотивирано от три основни дефицита:

***Дефицит 1 (D1): Липса на изследване, отчитащо прага на сходство*.** Прагът на сходство (similarity threshold) определя дали извлечените документи да бъдат включени в контекста за генериране на отговор и пряко влияе върху прецизността (precision) и пълнотата (recall) на извличането [6], [11]. На практика, изборът на този праг, не е просто технически детайл от реализацията: той определя дали системата ще подаде недостатъчен контекст (водещ до дефицит на доказателства и непълни отговори) или прекомерен такъв (въвеждащ ирелевантни пасажи, които могат да "разсеят" модела и да влошат фактологичното съответствие). Изборът на праг влияе и върху изчислителните

разходи, чрез контролиране на обема на контекста и последващата му обработка. Необходима е процедура за оценяване, която отчита прага на сходство. Тя позволява да се анализира селективността на извличане и да се обоснове изборът на конкретна конфигурация. Чрез вариране на прага при контролирани условия, се установяват характерни зависимости в резултатите, които са характерни за различните модели и набори от данни. Въпреки че прагът се използва масово като параметър, систематични анализи с последователното му вариране, при контролирани условия и съпоставими метрики, остават ограничени.

Бележка: *Под "праг на сходство" се разбира минималният коефициент на косинусово сходство (cosine similarity score), необходим за включването на даден пасаж в контекста за генериране. В литературата този параметър се среща още като "праг на релевантност" или "праг на селективност"; за улеснение в следващите глави ще се използва "праг на сходство" или просто "праг".*

Дефицит 2 (D2): Липса на инфраструктура за възпроизводимост. RAG системите включват множество взаимодействащи си конфигурационни слоеве: предварителна обработка на корпуса, стратегия за сегментиране (chunking), избор на модел за векторни вграждания (embeddings), индексиране, както и настройки за извличане, генериране и логика на оценяване. Поради комплексното взаимодействие между тези слоеве, независимата проверка е трудна, ако конфигурациите не са прецизно дефинирани и представени в пълна и съпоставима форма [12], [13]. Липсата на прецизна документация прави резултатите трудни за проверка. В такива случаи е невъзможно да се установи дали подобрената производителност се дължи на изследваната промяна или на скрити разлики в настройките на системата. Този проблем е особено критичен при изследвания, чувствителни към прага на сходство, където дори минимални промени в конфигурацията на извличане могат да изменят коренно доказателствата, подавани към модела. Ето защо е необходима инфраструктура, която прави експериментите проверими и сравними, документира и съхранява пълния контекст и резултатите от експерименталните цикли във формат, позволяващ независима верификация и сравнение между различните изследвания.

Дефицит 3 (D3): Липса на практически насоки за внедряване на модели с отворен код. Организациите, които поради съображения за сигурност или разходи предпочитат собствена инфраструктура (on-premises), разчитат предимно на модели с отворен код. Въпреки това, все още липсват задълбочени анализи, които да свързват прага на сходство с качеството на отговорите и ресурсите, необходими за поддръжка на системата. При липсата на такива данни, изборът на модел често се основава на неформални тестове или на грешни предположения за поведението на извличащите алгоритми [10], [14]. Необходими са практически насоки, които да илюстрират влиянието на прага върху работата на системата и да позволят обективно сравнение между различните модели с отворен код.

Цел на изследването

Целта на дисертационния труд е да се разработи рамка за оценка на RAG, която подпомага вземането на решения за конфигуриране на извличането в RAG системи при езикови модели с отворен код, с особен акцент върху настройването на прага на сходство.

Изследователски въпроси

- **Въпрос 1 (RQ1):** Води ли промяната на прага на сходство до измерими промени в качеството на генерираното съдържание?
- **Въпрос 2 (RQ2):** Различава ли се ефектът от прага на сходство при различните езикови модели?
- **Въпрос 3 (RQ3):** Валидни ли са сходни диапазони на прага при различна тематична област (домейн)?

Всеки изследователски въпрос е обвързан с конкретна експериментална фаза: **RQ1** се разглежда чрез последователна промяна на прага във Фази II-IV; **RQ2** - чрез сравнителен анализ на моделите във Фази III-IV; а **RQ3** - чрез съпоставяне на резултатите при използване на данни от различни тематични области - земеделие и биоразнообразие - във Фаза IV. Първата фаза (Фаза I) е насочена към демонстрация на системата и профилиране на производителността с цел да се установи базова функционалност преди началото на експериментите с прага на сходство.

Задачи

Определени са четири задачи:

Задача 1 (Obj.1): Да се дефинират и реализират основните компоненти на рамката за оценка, интегрираща три компонента: **(a)** процедура за оценка, отчитаща прага на сходство чрез композитни показатели (composite scoring); **(b)** платформата за оценка на производителността PaSSER (Performance Assessment System for Similarity Evaluation and Retrieval) [15], осигуряваща инфраструктура за възпроизводимост, чрез запис на резултатите върху блокчейн; и **(c)** контролиран експериментален дизайн, за извеждане на сравнителни данни между различни модели и области.

Задача 2 (Obj.2): Установяване на критерии за избор на модели. Дефиниране на критерии за подбор, съобразени с възможностите за локално внедряване, лицензионен режим и изчислителни ресурси. Това включва профилиране на избраните модели по отношение на техния контекстен прозорец и декодинг настройки.

Задача 3 (Obj.3): Определяне на процедури за подбор и изчисляване на метрики. Да се подберат метрики, съответстващи на оценяваните характеристики - лексикално припокриване, семантично сходство, гладкост на текста, точност и езиково моделиране - и да се осигури последователното им изчисляване при всички модели и експериментални условия.

Задача 4 (Obj.4): Провеждане на контролирано тестване и анализ. Подготовка на тематични масиви от данни (корпуси) и набори от въпроси и отговори, включваща специфична предварителна обработка и конфигурация на извличането. Изпълнение и оценка на серия от тестове и изследване на целия диапазон от прагове на сходство (threshold sweeps). Анализ и обобщаване на резултатите, за да се установи как прагът на сходство влияе върху качеството на генериране, възпроизводимостта и избора на подходяща конфигурация.

Връзка между изследователските въпроси и поставените задачи. Изследователските въпроси RQ1-RQ3 имат емпиричен характер и разглеждат как конфигурацията на прага на сходство влияе върху работата на RAG системи при различни модели и тематични области. Задача 1 определя основните компоненти на рамката за оценка, включително процедурата за оценка, инфраструктурата за възпроизводимост и контролиран експериментален дизайн. Задачи 2-4 конкретизират тази рамка чрез подбор на модели, определяне и изчисляване на метриките, както и чрез контролирано тестване и анализ. Заедно те осигуряват основата за отговор на RQ1-RQ3 и за извеждане на практически насоки при избора на модел и праг.

Таблица I.1 проследява връзката между идентифицираните дефицити, основните въпроси на изследването, съответните задачи и научни приноси.

Таблица I.1 *Съответствие на дефицити, основни въпроси, задачи и приноси.*

Дефицит	Изследователски въпроси	Задачи	Глави	Принос	Компонент от рамката
D1: Изследване, отчитащо зависимостите при вариране на прага	RQ1, RQ2, RQ3	Obj.1 (a), Obj.3, Obj.4	Гл. 3 - 4	C1	Процедура за оценка
D2: Инфраструктура за възпроизводимост	—	Obj.1 (b)	Гл. 2	C2	Инфраструктура
D3: Липса на практически насоки за избор на модел и праг	RQ1, RQ2, RQ3	Obj.1 - 4	Гл. 3 - 4	C3	Емпирични данни

* *Задача 1 допринася за преодоляването и на трите дефицита, като дефинира процедурата за оценка, отчитаща прага на сходство (D1), реализира проследяване на произхода и регистриране в блокчейн (D2) и установява контролиран експериментален дизайн, който осигурява сравнителни данни (D3). Задачи 2-4 представляват конкретни компоненти в обхвата на тази рамка.*

Три научно-приложни приноса представляват рамката за оценяване.

Принос 1 (C1): Компонент "Процедура за оценка". въвежда процедура за оценка, отчитаща прага на сходство, която включва Композитен показател за производителност (Composite Performance Score, CPS), Композитен показател, за стабилност на прага (Threshold-aware Composite Performance Score, T-CPS) и Показател за баланс (Balance Score, BS) за характеризиране на селективността на извличането при различни настройки на прага на сходство [15], [16].

Принос 2 (C2): Инфраструктурен компонент. Реализира се инфраструктура за тестване и възпроизводимост на резултатите, съчетаваща запис на експерименталните данни в блокчейн и пълно

документиране на конфигурациите чрез платформата **PaSSER** [15], [17].

Принос 3 (С3): Компонент "Емпирични данни". Извежда практически насоки за внедряване на RAG системи с модели с отворен код, основани на сравнителни емпирични данни, които свързват чувствителността към прага на сходство, качеството на генериране и приложимостта при внедряване при седем модела в диапазона 7-8 милиарда параметри, изследвани при контролирани експериментални условия [16], [18].

Структура

Дисертацията се състои от: увод, пет глави, заключение, приложения и библиография.

Първа глава поставя основите на дисертацията, като прави преглед на научните публикации по темата за RAG, практиките за оценка и предизвикателствата пред възпроизводимостта. Също позиционира дефицитите D1–D3 в рамките на съответната литература.

Втора глава представя инфраструктурата на платформата PaSSER. В нея се разглеждат работният процес, конфигурирането на различните настройки, автоматизираният процес на тестване и проследяването на произхода чрез блокчейн.

Трета глава обосновава избора на модели и определя метриките за оценка и процедурите за тяхното изчисляване, прилагани при сравнителната оценка между моделите.

Четвърта глава представя емпиричните резултати от контролирано тестване върху наборите от данни в областите земеделие и биоразнообразие, като анализира чувствителността към прага на сходство, представянето на моделите и съпоставката между тематичните области.

Пета глава обсъжда изследователските въпроси и научно-приложните приноси, разглежда ограниченията и очертава бъдещи насоки за изследване.

Заключението обобщава основните резултати.

Заключителната част включва публикации подкрепящи дисертацията, справка на забелязаните цитирания, обобщение на участието в проекти, благодарности и декларация за оригиналност на резултатите.

Забележка: *Таблиците и фигурите са запазени само в случаите, когато са необходими за разбирането на синтезираното съдържание. Номерацията на всички включени таблици и фигури е запазена така, както е в дисертационния труд, в съответствие с изискванията за автореферат.*

ПЪРВА ГЛАВА: ГЕНЕРИРАНЕ, ПОДПОМОГНАТО ЧРЕЗ ИЗВЛИЧАНЕ

Големите езикови модели генерират четим и смислен текст, но са ограничени от статично параметрично знание. Отговорите им могат да бъдат необосновани, остарели или трудни за проследяване. RAG смекчава тези ограничения, като включва външни доказателства по време на генерирането на отговор. Първа глава позиционира дисертационния труд, като проследява развитието на RAG от извличането на информация (Information Retrieval, IR) [11] и обработката на естествен език (Natural Language Processing, NLP) [19], представя съществуващите RAG архитектури и характерните им недостатъци, разглежда подходи за оценяване и идентифицира липсите в литературата, съответстващи на дефицити D1-D3.

1.1 Първоначално развитие

RAG стъпва върху дългогодишното развитие на IR и NLP и върху все по-тясната им връзка [20]. Секцията проследява този процес чрез четири направления, които оформят съвременните RAG системи: (1) базово индексиране и организация на данни, (2) формално оценяване в IR и ранни опити за съчетаване на IR и NLP, (3) развитие на семантично извличане и методи от NLP и (4) мащабна интеграция на IR и NLP с модерни подходи на основата на вграждания [20]. Ранните практики в индексирането и оценяването, въвеждат понятия за ефективност на извличането, като прецизност и изчерпателност [11], а семантичното извличане постепенно преминава към сравнение по сходство, чрез векторни представяния (embeddings). Това развитие дава техническия контекст за интегрирането на извличането на релевантни пасажи (от корпус) със съвременните генеративни модели.

1.2 Появата на RAG

Концепцията за RAG е представена през 2020 г. в "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" [5]. Тя съчетава извличането на информация от външни източници с

генерирането на текст, за да намали ограниченията на изцяло параметричните модели, включително зависимостта от статични данни, ограничените възможности за позоваване на източници и халюцинациите [21], [22]. Архитектурно RAG разделя извличането (retrieval) и генерирането (generation): при извеждането на отговор, релевантни пасажи се извличат от колекция от документи и се използват като контекст при генериране. Това подпомага фактическата точност, прозрачността и адаптацията към тематична област без преобучаване на модела. Извличането използва векторни представяния (например DPR [23] с BERT-базирани кодиращи модели [24]) и търсене по сходство (например MIPS [25]) с приблизителни методи за намиране на най-сходните пасажи [26], [27]. Генерирането е реализирано в схема "последователност-към-последователност" (например BART [28]), стъпвайки върху по-ранни парадигми за машинен превод [29]. В оригиналната формулировка RAG се обучава съвместно като единна система, така че извличането да е съобразено с нуждите на генерирането. Докладваните резултти обхващат отговаряне на въпроси в отворена област и други задачи, които изискват външни знания, включително Natural Questions [30], TriviaQA [31], FEVER [32] и MS MARCO NLG [33]. Резултатите показват конкурентна производителност с по-малко параметри спрямо изцяло параметрични подходи [5].

1.3 Иновации и разширения на RAG

Последващи изследвания разширяват RAG с цел по-добро извличане, по-точни отговори, по-лесна интерпретация и по-висока ефективност. Таблица 1.2 представя функционална класификация на тези разработки в седем направления [20]: архитектурна ефективност и мащабируемост, оптимизация на данните, итеративно извличане и самоусъвършенстване, интеграция на знания и мултимодални разширения, адаптация и специализация по тематични области, проверка на факти и обосноваване чрез източници, както и подобряване при малко примери и ограничени ресурси. Макар направленията да се различават по механизъм, много от тях разчитат на селективност при извличането, която определя какви доказателства достигат до генератора.

Таблица 1.2 *Функционално разпределение на съвременните подобрения в RAG системите. Източник [20]*

Фокус	Цел
Архитектурна ефективност и мащабируемост	Намаляване на изчислителните разходи, ускоряване на извеждането на отговор и поддръжка на приложения в реално време
Оптимизация на данните	Подобряване на качеството на обучението чрез намаляване на шума, семплиране и техники за подбор на данни
Итеративно извличане и самоусъвършенстване	Въвеждане на механизми за многостъпково разсъждение, редакция на отговора или извличане на база обратна връзка
Интеграция на знания и мултимодални разширения	Съчетаване на символно и невронно извличане за достъп до различни представяния на знанията; разширяване на извличането към допълнителни модалности
Адаптация и специализация по предметни области	Осигуряване на добро представяне на RAG системите в специализирани предметни области или тесни тематични направления
Проверка на факти и обосноваване чрез източници	Намаляване на халюцинациите и повишаване на прозрачността чрез обвързване на отговорите с проверими източници
Подобряване при малко примери и ограничени ресурси	Подобряване на обобщаването при ограничени обучителни данни чрез few-shot обучение, подпомогнато от извличане

Секции 1.3.1-1.3.7 от дисертацията разглеждат всяка категория в детайли; по-долу е обобщена подсекцията с най-пряко отношение към настоящото изследване.

1.3.8 Конфигурация на извличане и избор на праг на сходство

Селективността при извличането най-често се управлява чрез два механизма: *top-k* извличане и филтриране по праг. *Top-k* връща фиксиран брой пасажи с най-висок резултат, което може да включи и слабо релевантно съдържание, когато липсват достатъчно силни съвпадения. Филтрирането по праг връща само пасажи, чието сходство надвишава зададена минимална стойност. Така размерът на контекста става променлив, а при слаби доказателства е възможно да не бъде върнат нито един пасаж. На практика често се използва комбинация: първо се прилага праг, а след това резултатите се

ограничават до *top-k* (или се задава минимална стойност за сходство в рамките на *top-k*). Това създава набор от конфигурации, който пряко влияе върху точността и изчерпателността на извличането.

В системите, представени в литературата, прагови механизми се въвеждат чрез правила за насочване, тригери за увереност или граници за вземане на решение (например Self-RAG [34], CRAG [35], Adaptive-RAG [36], FLARE [37]). Праговете на сходство обикновено се избират чрез локална настройка или практически правила, а не чрез систематичен анализ на чувствителността. Стойностите на прага не се пренасят надеждно между различни модели за векторни представяния, функции за сходство и корпуси. Праг, който работи добре в една конфигурация, може да доведе до масово допускане на нерелевантни пасажи в друга. В индустриален казус от банковата област, при баова конфигурация с фиксиран праг на сходство 0.7, е отчетен дял на фалшивите положителни резултати до 99% при някои модели за векторни представяния. [38]. Повечето изследвания докладват резултати за една конфигурация на извличане и не показват как се променя производителността при вариране на прага. Това ограничава избора на праг, основан на доказателства, затруднява сравненията между модели при различни режими на селективност и намалява преносимостта на изводите между тематичните области.

1.4 Оценяване на RAG

Оценяването на RAG изисква повече от проверка на точността на отговора или на лексикално припокриване [39]. Важно е и дали извлечените доказателства са релевантни, както и дали генерираният отговор се опира на тях [40], [14]. Подходи като RAGAS [40], RGB [14] и TREC 2024 RAG Track [41] предлагат набори от метрики и споделена инфраструктура. Инструменти като TruLens [42] и други платформи за проследяване и наблюдение (например, LangSmith [43], [44] и Arize Phoenix [45]) подпомагат диагностика на ниво изпълнение и сравнение, подобни на регресионно тестване. Тестови рамки като DeepEval [46] интегрират автоматизирано оценяване в в работния поток.

Секции 1.4.1-1.4.5 от дисертацията разглеждат всеки подход в детайли; установените липси в литературата са обобщени по-долу.

1.4.6 Ограничения в оценяването на RAG

В разгледаните подходи се открояват три ограничения.

Първо, резултатите най-често се докладват при една фиксирана конфигурация на извличане, без анализ на поведението при различни стойности на прага на сходство (D1).

Второ, възпроизводимостта често страда от непълно записване на конфигурацията, а логовете от проследяване, сами по себе си, не осигуряват надеждна връзка между набора от данни, настройките, междинните артефакти и получените резултати (D2).

Трето, сравнителните резултати често се представят спрямо затворени базови решения, а насоките за внедряване с отворен код при реални ограничения остават ограничени или разпокъсани (D3).

Тези ограничения мотивират необходимостта от Рамка за оценка. Ограниченията, свързани с фиксирани конфигурации на извличането, съответстват на D1 и се разглеждат в компонента "Процедура за оценка". Ограниченията, свързани с документирането на конфигурацията и проследяването на произхода, съответстват на D2 и се разглеждат в инфраструктурния компонент. Ограниченията, свързани със сравнителните практически насоки при ограниченията на внедряването с модели с отворен код, съответстват на D3 и се разглеждат в компонент "Емпирични данни".

1.5 Текущи предизвикателства и нови решения

Повтарящите се проблеми при RAG включват: липсващо съдържание, пропускане на най-високо класирани документи, фрагментиран контекст, некачествено извличане на съдържание, непоследователно структуриране на отговора, неподходяща степен на конкретност и непълни отговори [47]. Скорошните разработки се опитват да ограничат тези проблеми (напр. Auto-RAG [48], GraphRAG [49], семплиране по релевантност [6], FLARE [37], LightRAG [50], Self-RAG [34], Speculative RAG [51]), но ефективността им зависи от оценяване, което систематично открива типичните проблеми и показва влиянието на селективността.

1.6 Първа глава: Обобщение

Първа глава представя основите на RAG, разглежда различните направления на развитие и подходите за оценяване, като посочви селективността при извличането - в частност избора на праг на сходство - като фактор с недостатъчно изследвано влияние върху качеството на генерирането.

Формулирани са три ограничения в литературата, съответстващи на D1-D3: ограничени доказателства за оценяване с отчитане на прага, непълна среда за възпроизводимост която да надхвърля записването на логове, и ограничени насоки за избор и имплементация при използване на модели с отворен код. Тези ограничения обосновават инфраструктурата, описана във Втора глава, процедурите за избор на модели и метрики в Трета глава и контролираните експерименти с последователно вариране на прага, докладвани в Четвърта глава.

ВТОРА ГЛАВА: ДИЗАЙН И АРХИТЕКТУРА НА PaSSER

PaSSER (Performance Assessment System for Similarity Evaluation and Retrieval) е модулна уеб-базирана платформа за конфигуриране и оценяване на RAG при големи езикови модели с отворен код [15], [16], [17]. Платформата обединява извличане с регулируем праг на сходство, оценяване по набор от показатели и запис на резултатите чрез блокчейн, в единен работен процес за контролирани експерименти. Така се адресира D2 (инфраструктура за възпроизводимост) и се реализира инфраструктурният компонент (б) на Obj.1. Втора глава представя дизайна, архитектурата на PaSSER и работния процес за автоматизирано оценяване, използван в контролираните експерименти, описани в Четвърта глава.

2.1 Първоначален проект на системата

PaSSER е разработен като допълнителен модул към платформата Smart Crop Production Data Exchange (SCPDx) [52], [53], [54]. SCPDx комбинира блокчейн компонент, базиран на Antelope (предишно наименование EOSIO) [55], с InterPlanetary File System (IPFS) [56] за децентрализирано управление на данни. Предходните изследвания обосновават избора на блокчейн платформата, включително оценка на приложимостта ѝ [52] и интеграцията на оракули [57]. Блокчейнът Antelope е избран поради възможността за одит, устойчивостта на записите срещу промени, ясния модел за права на достъп и ниската цена на транзакциите, подходяща при чести записи. Anchor Wallet [58], [59] осигурява подписването на транзакциите от страна на клиента, като частните ключове не се използват в уеб приложението.

Въпреки че SCPDx интегрира IPFS за разпределено съхранение, PaSSER понастоящем не използва IPFS; бъдещата интеграция е отбелязана в Секция 5.4.2. PaSSER е проектирана като уеб-базирана среда с цел независимост от конкретна платформа и минимални изисквания за локална инсталация, което подпомага възпроизводимостта при различни клиентски устройства.

2.2 Системна архитектура

PaSSER използва трислойна архитектура: уеб интерфейс (SPA), бекенд услуги и блокчейн подсистема. Уеб интерфейсът служи за въвеждане на конфигурации и визуализация на резултатите, а изчисленията се изпълняват от бекенда, така че извличането, генерирането и оценяването да протичат по един и същи начин, независимо от клиентската среда. Автентикацията и подписването на транзакции се извършват от клиентската страна чрез Anchor Wallet, а блокчейн подсистемата записва данни за последваща проверка след приключване на изпълнението [17].

Секция 2.2.1 в дисертацията описва уеб интерфейса в детайли; в автореферата акцентът е върху бекенд процеса и блокчейн подсистемата, като основни компоненти в тестването и оценяването, и техният последващ запис, защитен срещу промени.

2.2.2 Бекенд услуги

Бекенд услугите изпълняват основния процес на PaSSER: извличане, генериране на отговор чрез езиков модел, организиране на оценяването и осигуряване на проследимост. Семантичното извличане е реализирано с ChromaDB [60]. Корпусите от документи се преобразуват във векторни представяния чрез API на Ollama и се съхраняват като векторни колекции. При добавяне на документи текстът се разделя на припокриващи се фрагменти; PaSSER позволява да се задават размерът на фрагмента и припокриването като конфигурационни параметри. Стандартната конфигурация е 1024 знака с припокриване от 50 знака и е използвана в контролираните експерименти в Глава 4. Фрагментирането влияе върху детайлността на извличането и свързаността на контекста и може да се отрази на селективността [61]; ефектите от чувствителността към начина на фрагментиране са разгледани в Секция 5.3.2.

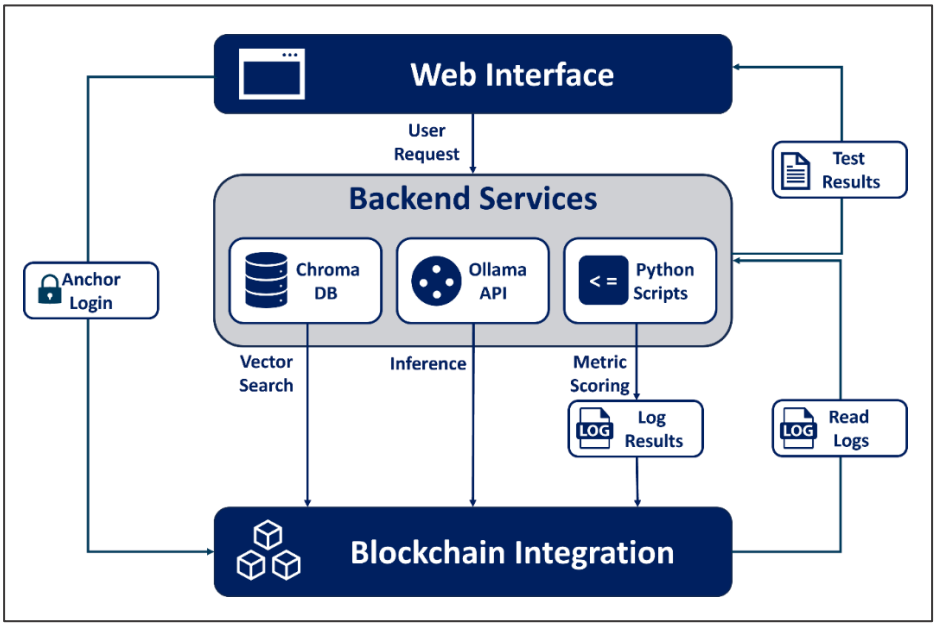
Режими на извличане. PaSSER работи в два режима. "Normal Mode" извиква VectorStoreRetriever от LangChain [62] и връща top-k пасажи, подредени по косинусово сходство, като

броят на пасажите е фиксиран независимо от абсолютните стойности на сходство. "Score Mode" използва ScoreThresholdRetriever за филтриране на пасажите по минимална стойност на косинусово сходство. "Score Mode" предоставя три параметъра: minSimilarityScore (минимална стойност за включване), maxK (горна граница на върнатите пасажи) и kIncrement (стъпка за итеративно вариране на прага). "Normal Mode" осигурява базовата линия с фиксиран top-k (Фаза I, Секция 4.1), а "Score Mode" поддържа контролираните експерименти за чувствителност към прага (Фаза II, Секция 4.2).

Генериране и оценяване. Генерирането на текст се извършва през Ollama API (приложен интерфейс) [63]. Получените отговори се оценяват спрямо референтни отговори от Python скрипт за оценяване, който използва утвърдени библиотеки: Natural Language Toolkit (NLTK), torch, NumPy, rouge, transformers и SciPy [17]. Формалните дефиниции на метриките са представени в Трета глава, а в Четвърта глава са уточнени използваните метрики и композитните показатели (включително CPS и T-CPS) за всяка експериментална фаза.

Осигуряване на проследимост. При всяко изпълнение се записват активната конфигурация (идентификатор на модела, параметри на извличане, настройки за декодиране, идентификатори на набора от данни и векторното хранилище), както и времеви и качествени показатели. Кратки обобщения се подават към блокчейна Antelope чрез конектор, реализиран с Pyntelope [64]. Подписването на транзакциите се извършва отделно чрез Anchor Wallet, а подаването - от бекенд услугата.

Фигура 2.2 обобщава работния процес: извличане с ChromaDB, генериране с Ollama, оценяване с Python и записване в блокчейна.



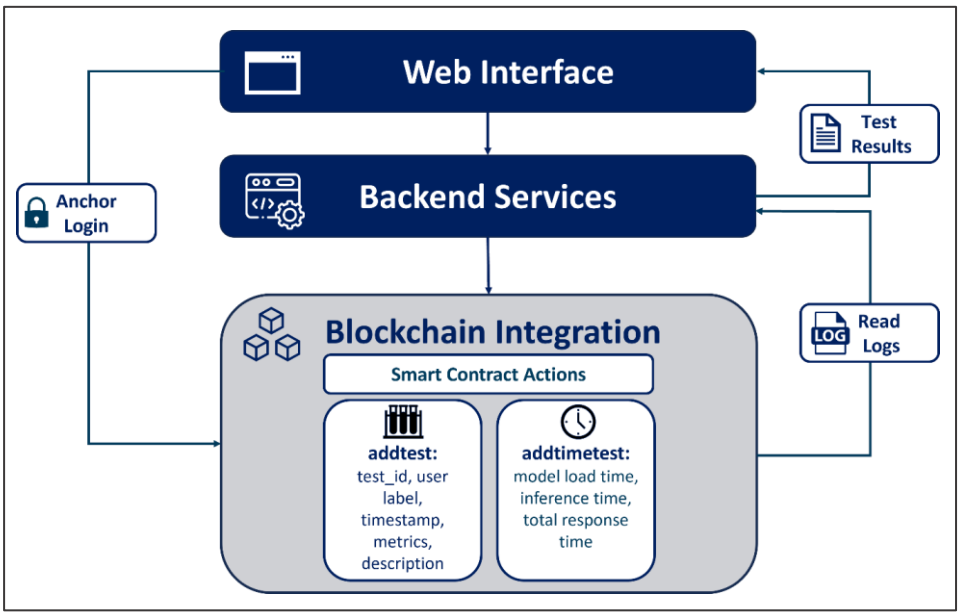
Фигура 2.2 *Бекенд услуги в PaSSER. Източник [16].*

2.2.3 Интеграция с блокчейн

PaSSER използва блокчейна Antelope (по-рано EOSIO) за запис на резултатите от оценяването и съпътстващите метаданни по начин, който не допуска последващи промени. Блокчейнът не е част от пътя на извличане и генериране; записването се извършва след приключване на изпълнението и служи за последваща проверка. Автентикацията и подписването на транзакциите се извършват чрез Anchor Wallet, а бекенд услугата подготвя данните и подава подписаните транзакции, без достъп до частните ключове.

Резултатите от оценяването се записват чрез смарт договор "*llmtest*" с две действия, които работят само в режим "добавяне": "*addtest*" записва качествените метрики, а "*addtimetest*" - времеви показатели (например време за зареждане на модела и продължителност на генерирането). Независима проверка може да се извърши чрез извличане на съответния запис от блокчейна и сравняване на съхранените полета (акаунтът, който подава записа, идентификатор на изпълнение, времева марка, описателен етикет и числовият масив, съхранен като float64[]) с експортираните артефакти от изпълнението, използвани при анализа.

Фигура 2.3 показва действията "*addtest*" и "*addtimetest*" и етапа, в който в блокчейна се записват обобщени стойности и метаданни за изпълнението.



Фигура 2.3 *Блокчейн интеграция в PaSSER. Източник [16].*

2.3 Функционалности на PaSSER

Тази секция обобщава потребителския работен процес и артефактите, необходими за контролирани експериментални цикли. Платформата поддържа както интерактивна работа, така и автоматизирано оценяване - изпълнение на серия от тестови въпроси без ръчна намеса, при което системата генерира отговори, изчислява метрики и записва резултатите. Експериментите, описани в Глава 4, са проведени с автоматизирано оценяване.

2.3.1 Конфигуриране на системата

Конфигурирането включва задаване на крайни точки за сървъра за генериране (Ollama API) и векторното хранилище (ChromaDB), избор на активен езиков модел и температура на генериране, избор на режим на извличане (Normal или Score Mode), както и свързване на ресурси за данни (избор на векторна колекция и импорт на JSON набор от двойки въпрос-отговор). След потвърждение конфигурацията е обвързана със сесията и се прикрепя към изходните данни за проследимост и възпроизводимост.

Секции 2.3.2 и 2.3.4 в дисертацията описват допълнителни работни процеси на ниво интерфейс; авторефератът обобщава единствено конфигурирането, управлението на извличането и автоматизираното тестване и оценяване, използвани в Четвърта глава.

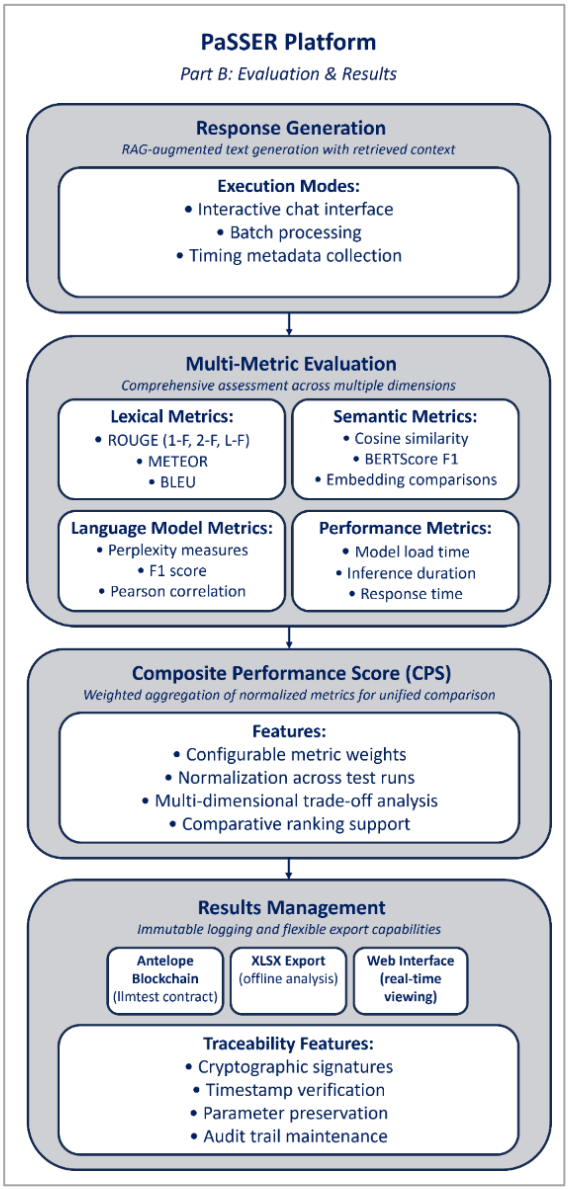
2.3.3 Конфигуриране на извличането

Конфигурацията на извличането определя как пасажите се подбират и сглобяват като контекст. **Normal Mode** осигурява фиксирано *top-k* извличане. **Score Mode** позволява филтриране по праг чрез минимален коефициент на косинусово сходство и поддържа контролирано вариране на прага чрез параметрите *minSimilarityScore*, *maxK* и *kIncrement*. Варирането на прага (threshold sweep) се реализира чрез повторно изпълнение на едно и също тестово натоварване при промяна единствено на стойността на прага, като всички останали параметри се поддържат постоянни (набор от данни, идентификатор на векторното хранилище, конфигурация на фрагментирането и настройки на модела). Всяко изпълнение се обозначава и записва като отделен запис за оценяване, което позволява пряко сравнение между различните стойности на прага при фиксирана конфигурация.

2.3.5 Тестване и оценка

PaSSER поддържа автоматизирано тестване и оценяване, на набор от данни, при контролирани конфигурации на извличане. За всеки елемент от набора системата извлича контекст, конструира обогатена заявка (augmented prompt), генерира отговор, изчислява метрики за оценяване и записва резултатите в блокчейна чрез *addtest*. Допълнителен модул за времево оценяване регистрира показатели за латентност (например време за зареждане на модела, продължителност на

генерирането, общо време за отговор) чрез *addtimetest*. Резултатите могат да бъдат извлечени за преглед и експортирани за офлайн анализ. Фигура 2.7 показва логиката на оценяване.



Фигура 2.7 *Обобщен преглед на работния процес за оценяване в PaSSER. Източник [16].*

2.4 Втора глава: Обобщение

Глава 2 представя PaSSER като инфраструктура за тестване и оценяване на RAG с отчитане на прага на сходство [15], [16], [17]. Платформата интегрира бекенд услуги, координиращи извличането и генерирането, с блокчейн подсистема Antelope, осигуряваща проследим и устойчив на промени запис на данните. PaSSER поддържа както фиксирано top-k извличане, така и извличане с филтриране по праг, което позволява контролирано вариране на прага при постоянни експериментални условия. Към всяко изпълнение се добавят конфигурационни метаданни, а в блокчейна се записват подписани записи чрез "*addtest*" и "*addtimetest*". Това позволява одит и независима проверка на отчетените резултати, адресира D2 и подпомага експериментите, описани в Четвърта глава.

ТРЕТА ГЛАВА: ИЗБОР НА МОДЕЛИ И МЕТРИКИ ЗА ОЦЕНЯВАНЕ

Трета глава определя компонентите, необходими за експерименталния анализ в Четвърта глава: големите езикови модели с отворен код, оценявани в PaSSER, и метриките за оценка на качеството на генерираните отговори. Секции 3.1-3.3 описват групите модели, аргументите за избора им и ограниченията при интеграция (Obj.2). Секция 3.4 представя метриките за оценяване, включително описание, начин на изчисляване и формат за отчитане. Секция 3.5 дефинира Композитен

показател за производителност (CPS) и Композитен показател, за стабилност на прага (T-CPS), използвани за агрегиране на набор от метрики (Задача 3). За краткост по-нататък "Композитен показател за производителност" се обозначава като CPS, а "Композитен показател, за стабилност на прага" – като T-CPS. Заедно тези елементи подпомагат адресирането на D1 (оценка с отчитане на прага чрез CPS, T-CPS и Показател за баланс (Balance Score)) и D3 (практически насоки за внедряване на решения с отворен код чрез критерии за избор на модели и процедури за оценяване).

3.1 Преглед на интегрираните модели и критериите за техния подбор

В PaSSER е интегрирана представителна група от езикови модели с отворен код, за сравнения при еднакви експериментални условия. Подборът е насочен към публично достъпни модели, възможност за локално изпълнение на хардуер от среден клас и архитектурно разнообразие в диапазона 7-8 милиарда параметри. Този диапазон балансира възможности и практическа приложимост, и обикновено позволява изпълнение при 16-32 GB оперативна памет (RAM), без отделни GPU системи, като същевременно избягва изискванията на по-големи модели от клас 40B. Модели под 7B не са включени поради ограничен капацитет при генериране в области, които изискват по-специализирани знания, а модели над 8B - поради по-високи хардуерни изисквания [65]. Не са включени също затворени (proprietary) модели, тъй като лицензите, цената и ограничената прозрачност на настройките правят трудно точното възпроизвеждане на конфигурацията [66].

Оценени са две групи модели: първа група от 7B модели, използвана във Фази I-II, и втора група (клас 8B, плюс обновен Mistral), използвана във Фази III-IV. Архитектурното разнообразие е важно, тъй като чувствителността към прага може да зависи от особености на конкретния модел, например ограничения на контекстния прозорец, ефективност на механизма за внимание и настройка за следване на инструкции или разсъждение [5], [67], [68]. Таблица 3.1 обобщава основните характеристики на оценяваните модели.

Таблица 3.1 *Сравнително обобщение на оценяваните модели.*

Модел	Парам.	Контекст (токени)	Основен акцент в дизайна	Основна роля в оценяването	Ollama таг
Mistral 7B / v0.3	~7.3B	8,192 / 32,768	GQA (grouped-query attention); внимание с плъзгащ прозорец	Фокус върху ефективност; приемственост между версии	Mistral 7B / Mistral:latest
Llama 2 7B	7B	4,096	Стандартен трансформър; широко използван	Базова линия за общо предназначение	Llama 2 7B
Orca 2 7B	7B	4,096	Дообучаване с фокус върху разсъждение	Сравнителен модел, настроен за разсъждение	Orca 2 7B
Granite 3.2 8B	8B	8,192	Подбор и подготовка на данни с корпоративен фокус	Сравнителен модел за надеждност в корпоративни среди	Granite3.2:8b
DeepSeek R1 8B	8B	8,192	Обучение с подсилване (RL) с фокус върху разсъждение	Сравнителен 8B модел, фокусиран върху разсъждение	Deepseek r1:8b
Llama 3.1 8B	8B	8,192	Обновен LLaMA; разширен контекст	Базова линия от текущото поколение	llama3.1:8b

3.2 Начална група модели

Началната група - Mistral 7B [69], Llama 2 7B [70] и Orca 2 7B [71] - е използвана за проверка на цялостната работа на системата и за измерване на времевите показатели при фиксирано top-k извличане във Фаза I [17]. Във Фаза II [15] същите модели са оценени при систематично вариране на прага (0,50-0,80), за да се опише чувствителността към прага.

3.2.1 Mistral 7B

Mistral 7B (2023) е проектиран за по-ефективно генериране в класа 7B параметри [69], [72]. Моделът използва grouped-query attention (GQA) and sliding-window attention (SWA) - механизми, които

са релевантни при RAG, тъй като добавянето на извлечени пасажи увеличава дължината на заявката и натоварва паметта и времето за изпълнение. Според резултатите, докладвани от авторите, Mistral 7B превъзхожда Llama 2 7B при стандартни оценки [73], включително MMLU [74] и GSM8K [75]. В началната група Mistral 7B е включен като представител с акцент върху ефективността.

3.2.2 Llama 2 7B

Llama 2 7B (2023) е широко използван модел от клас 7B [70]. В началната група Llama 2 7B е включен като базова линия за сравнение, тъй като е добре познат, широко използван и позволява ефектите от настройките на извличането (top-k спрямо филтриране по праг) да се разглеждат спрямо стабилен ориентир при еднакви експериментални условия. Meta публикува и дообучени варианти за следване на инструкции [68], [70].

3.2.3 Orca 2 7B

Orca 2 7B (2023) е модел, базиран на Llama 2, настроен да подобрява разсъждението чрез подбрани инструкции, без промени в архитектурата [71]. Докладваните подобрения се в задачи за сравнително оценяване на разсъждение, като GSM8K и BIG-bench Hard [76]. В началната група Orca 2 7B е включен като представител с акцент върху разсъждението.

3.3 Актуализирана група модели

Актуализираната група - Granite 3.2 8B [77], DeepSeek R1 8B [78], Llama 3.1 8B [79] и Mistral 7B v0.3 [80] - е включена във Фаза III (анализ на чувствителността към прага за всеки модел поотделно) и във Фаза IV (анализ между тематичните области). Mistral е запазен като свързващ модел между фазите, но е обновен до v0.3.

3.3.1 Granite 3.2 8B

Granite 3.2 8B (IBM, 2024) насочен към корпоративна употреба и разработен за задачи като отговаряне на въпроси, обобщаване, извличане на информация, програмиране и RAG [77]. В актуализираната група, моделът е включен като представител, ориентиран към корпоративни сценарии, при които се търси по-предвидимо поведение и стабилност при интеграция.

3.3.2 DeepSeek R1 8B

DeepSeek R1 8B (2024) поставя акцент върху обучителни подходи, ориентирани към разсъждение, и включва компоненти за обучение с подсилване [81]. Подходите за разсъждение, развити в по-големите варианти от серията DeepSeek R1, са пренесени към по-малкия R1 8B [78]. В актуализираната група изпълнява ролята на модел с акцент върху разсъждението.

3.3.3 Llama 3.1 8B

Llama 3.1 8B (Meta, юли 2024) е многоезичен модел, донастроен за следване на инструкции, от семейството Llama 3.1. Meta представя варианта 8B като модел с общо предназначение за асистентски задачи и посочва контекстен прозорец от 128k токена за текстовите модели от серията Llama 3.1. [79]. В актуализираната група служи като базова линия с общо предназначение при мащаб [82].

3.3.4 Mistral 7B (v0.3)

Mistral 7B v0.3 (2024) се разпространява чрез Ollama с идентификатор mistral:latest [63], предлага контекстен прозорец от 32 768 токена и извикване на функции [80], като запазва архитектурата, насочена към ефективност, описана в Секция 3.2.1. Моделът е включен за приемственост между фазите и за отчитане на промените в рамките на семейството Mistral [69], [72].

3.4 Метрики за оценяване

Оценяването на RAG изисква оценка в няколко направления, тъй като извличането пряко влияе върху генерирането. Използван е панел от 24 метрики, които обхващат лексикално припокриване, семантично съответствие, плавност и предсказуемост, качество на отговора, статистическа зависимост и показатели за четимост. Подборът следва принципа, че разнообразни показатели дават по-надеждна обобщена оценка от която и да е единична метрика [83]. Шестнадесет метрики са въведени във Фаза I, а още осем са добавени във Фаза III. Таблица 3.2 изброява пълния панел от 24 метрики (категории, изходни колони при получаване на резултите и фаза на въвеждане).

***Таблица 3.2** Пълен списък на 24-те метрики за оценяване.*

Категория	Метрика	Изходна колона	Описание	Фаза
Лексикално припокриване	METEOR	METEOR	Подравняване на ниво токен със стеминг и синонимни съвпадения	Фаза I
	ROUGE-1	Rouge-1.r	Еднограмно припокриване (recall)	Фаза I
	ROUGE-1	Rouge-1.p	Еднограмно припокриване (precision)	Фаза I
	ROUGE-1	Rouge-1.f	Еднограмно припокриване (F1)	Фаза I
	ROUGE-2	Rouge-2.r	Биграмно припокриване (recall)	Фаза I
	ROUGE-2	Rouge-2.p	Биграмно припокриване (precision)	Фаза I
	ROUGE-2	Rouge-2.f	Биграмно припокриване (F1)	Фаза I
	ROUGE-L	Rouge-l.r	Най-дълга обща подпоследователност (recall)	Фаза I
	ROUGE-L	Rouge-l.p	Най-дълга обща подпоследователност (precision)	Фаза I
	ROUGE-L	Rouge-l.f	Най-дълга обща подпоследователност (F1)	Фаза I
	BLEU	BLEU	n-грамна точност с наказание за краткост	Фаза I
	F1 Score	F1 Score	F1 на ниво токен; диагностика за коректност на отговора	Фаза I
Семантично сходство	Cosine similarity	Cosine similarity	Сходство на вграждания между генериран и референтен текст	Фаза I
	BERTScore	Bert-Score.precision	Контекстуално токенно сходство (precision)	Фаза III
	BERTScore	Bert-Score.recall	Контекстуално токенно сходство (recall)	Фаза III
	BERTScore	Bert-Score.f1	Контекстуално токенно сходство (F1)	Фаза III
Плавност / Предсказуемост	Laplace perplexity	Laplace Perplexity	Предсказуемост при Laplace-изгладен биграмен модел	Фаза I
	Lidstone perplexity	Lidstone Perplexity	Предсказуемост при Lidstone-изгладен триграмен модел	Фаза I
Статист. корелация	Pearson correlation	Pearson correlation	Линейна зависимост между генерирани и референтни представяния	Фаза I
Четимост (B-RT)	B-RT Coherence	B-RT.coherence	Тематична фокусираност и локална организация	Фаза III
	B-RT Consistency	B-RT.consistency	Вътрешна непротиворечивост между твърденията	Фаза III
	B-RT Fluency	B-RT.fluency	Четимост и граматична свързаност	Фаза III
	B-RT Relevance	B-RT.relevance	Съответствие с формулировката на въпроса	Фаза III
	B-RT Average	B-RT.average	Средно аритметично на B-RT компонентите	Фаза III

Секции 3.4.1-3.4.5 в дисертацията съдържат пълните дефиниции, начините за изчисляване и пълно представяне на всяка метрика. Съответстващите формули (3.1)-(3.24) са включени в дисертацията и са представени също в [17], [15] и [18]. В автореферата присъстват само кратки описания за Секции 3.4.1-3.4.4 и основната формула за B-RT в Секция 3.4.5.

3.4.1 Метрики за лексикално припокриване

Метриките за лексикално припокриване сравняват генерираните отговори с референтните на ниво токени и n-грам. METEOR [84], ROUGE [85] и BLEU [86] дават допълващи се показатели за съвпадение на текста.

3.4.2 Метрики за семантично сходство

Метриките за семантично сходство оценяват близостта по смисъл, отвъд повърхностното съвпадение. Cosine similarity [87] и BERTScore [88] сравняват генерирани и референтни отговори чрез векторни представяния.

3.4.3 Метрики за плавност, предсказуемост и качество на отговора

Показателите за плавност и предсказуемост се изчисляват чрез perplexity (мярка за предсказуемост), реализирана с NLTK, като индикатори, независими от конкретния езиков модел.; евентуалните проблеми са обсъдени в Секция 5.3.3. F1 на ниво токен се използва като показател за качество на отговора, извлечен от припокриването между генерираното и референтното съдържание.

3.4.4 Метрики за статистическа корелация

Статистическата зависимост се оценява чрез корелация на Pearson за описване на линейни зависимости между избрани метрични стойности.

3.4.5 Метрика, базирана на човешка четимост (B-RT)

Показателят B-RT е регресионен индикатор, изведен по модела на Nubia, който приближава човешките оценки за четимост при RAG [89], [90]. Стойностите се използват като автоматизирани сравнителни метрики, а не като заместител на валидирани човешки преценки. Базовото сходство е дефинирано като:

$$s = cos\big(E_{cls}(Reference),\ E_{cls}(G)\big), \quad \textbf{(3.25)}$$

а агрегираният индекс:

$$B - RT.Average = \frac{Coherence+Consistency+Fluency+Relevance}{4} \quad \textbf{(3.26)}$$

3.5 Композитни показатели

Тъй като отделните метрики обхващат различни аспекти на качеството, за систематично сравнение е необходимо тяхното обобщаване. Въпреки че PaSSER изчислява и 24-те метрики, при формирането на CPS се използват девет от тях, за да се избегне двойно претегляне на силно корелиращи се метрики. Метричните семейства от Секция 3.4 са групирани в четири оценъчни измерения. Таблица 3.3 представя това съответствие и начина, по който се извършва обобщаването.

Таблица 3.3 *Съответствие между метричните семейства от Секция 3.4 и измеренията на CPS.*

Семейство метрики (Секция 3.4)	Основно предназначение	Измерение на CPS	Бележка
3.4.1 Лексикално припокриване	Повърхностно съвпадение с референтния отговор	Лексикално припокриване	Директно съответствие
3.4.2 Семантично сходство	Близост по смисъл отвъд точното съвпадение	Семантично сходство и подравняване	Директно съответствие
3.4.3 Плавност, предсказуемост, качество	Езикова предсказуемост и коректност на отговора	Разделяне: Езиково моделиране (perplexity); Плавност и коректност (F1)	Perplexity → Езиково моделиране; F1 → Плавност и коректност
3.4.4 Статистическа корелация (Pearson)	Линейна зависимост между метрики	Семантично сходство и подравняване	Групирана като индикатор за подравняване
3.4.5 B-RT (набор от метрики)	Многоаспектни метрики за четимост и качество	Разпределени между компонентите на измеренията	Всеки компонент е съотнесен към съответното измерение

3.5.1 Формулировка на CPS

CPS обобщава нормализирани метрични стойности чрез претеглена сума и е разработен с оглед на три основни затруднения при сравнението между метрики: различни скали, различна полярност и различна диагностична стойност [15].

За даден въпрос *q*, оценен с модел *m* при праг на сходство *t*:

$$CPSq = \sum_{i=1}^{n} w_i \times \left[d_i \frac{(m_{i,q} - min_i)}{(max_i - min_i)} + \frac{(1 - d_i)}{2}\right] \quad \textbf{(3.27)}$$

където m_{i,q} е стойността на метрика i за въпрос q; min_i и max_i са наблюдаваните крайни стойности в оценяваните данни; d_i ∈ {−1, +1} е индикатор за полярност (d_i = +1 ако по-висока стойност е по-добра; d_i = −1 ако по-ниска стойност е по-добра); w_i е теглото на метриката (Σw_i = 1).

За метрики с положителна полярност (d_i = +1):

$$Normalized\ value = \frac{m_{i,q} - min_i}{max_i - min_i} \quad \textbf{(3.28)}$$

За метрики с отрицателна полярност (d_i = −1) нормализацията е обърната:

$$Normalized\ value = \frac{max_i - m_{i,q}}{max_i - min_i} \quad \textbf{(3.29)}$$

Така всички нормализирани стойности попадат в интервала [0, 1], като по-високите стойности последователно показват по-добро качество, независимо от оригиналната полярност.

Средна стойност на CPS за модел m при праг t за Q въпроса:

$$\mu_{m,t} = \frac{1}{Q} \sum_{q=1}^{Q} CPS_q^{(m,t)} \quad \textbf{(3.30)}$$

3.5.2 Композитен показател, за стабилност на прага (T-CPS)

CPS отразява средната производителност, но не показва доколко резултатите са устойчиви между отделните въпроси. Конфигурация с висок среден CPS, но с голяма вариация, може да е по-малко надеждна от конфигурация с малко по-нисък среден резултат, но със стабилни отговори. T-CPS добавя механизъм "награда-наказание", базиран на коефициента на вариация (CV) [91].

$$CV_{m,t} = \frac{\sigma_{m,t}}{\mu_{m,t}} \quad \textbf{(3.31)}$$

Награда за стабилност и наказание за вариация:

$$T - CPS = \mu \times (1 + \alpha \times (1 - CV)) - \beta \times CV^2 \quad \textbf{(3.32)}$$

Компонентът за награда (1 + α × (1 - CV)) повишава резултата при ниска вариация. Компонентът за наказание β × CV² намалява резултата квадратично при нарастване на CV. Параметрите α = 0,1 и β = 0,05 въвеждат асиметрия 2:1: наградата за стабилност може да добави до около +10% при ниска вариация, а наказанието намалява резултата с до около -5% при максималния наблюдаван CV. Тези стойности не се представят като оптимални; те са начална настройка, съобразена с вариацията на CPS, наблюдавана в предварителни изпълнения.

Анализ на чувствителността при 25 комбинации от параметри (α ∈ {0,05; 0,10; 0,15; 0,20; 0,25} × β ∈ {0,025; 0,05; 0,075; 0,10; 0,15}) потвърждава стабилност на класирането: 29 от 31 конфигурации (93,5%) не показват промяна в ранга при нито една комбинация. α обяснява 99,87% от вариацията на T-CPS, което показва, че наградата за стабилност е основният фактор, а наказанието за вариация има вторична, коригираща роля [18].

3.5.3 Статистическа значимост

За да се провери дали разликите между конфигурациите с различен праг и базовата линия са статистически значими, се използва t-тестове по двойки, като всеки въпрос се сравнява със себе си при базовата линия и при съответния праг (α = 0,05). Така се ограничава влиянието на различията между самите въпроси. Значимостта се отбелязва по стандартен начин: * за p < 0,05; ** за p < 0,01; *** за p < 0,001.

Размерът на ефекта се оценява чрез d на Cohen за зависими извадки:

$$d = \frac{M_{diff}}{SD_{diff}} \quad \textbf{(3.33)}$$

където M_diff е средната стойност на разликите в CPS по въпроси (праг минус базова линия), а SD_diff е стандартното отклонение на тези разлики. Размерите на ефекта следват стандартните конвенции [92]: незначителен (d < 0,2), малък (0,2 ≤ d < 0,5), среден (0,5 ≤ d < 0,8) или голям (d ≥ 0,8).

Докладваните p-стойности не са коригирани за множествени сравнения. Всяка фаза включва 40 сравнения (4 модела × 10 прагови стойности); при α = 0,05 приблизително два значими резултата на фаза биха могли да се появят случайно при нулева хипотеза. Поради това значимостта се тълкува на ниво обща картина - последователност между праговете в рамките на модела, съгласуваност между фазите и съответствие с размерите на ефекта - а не като окончателно доказателство от единично сравнение (виж Секция 5.3.3).

3.5.4 Показател за баланс (Balance Score)

T-CPS включва стабилността в композитния показател, но на практика, често е нужен отделен

критерий, който ясно да показва баланса между подобрение и нестабилност на резултатите. Balance Score измерва какво подобрение, коригирано за стабилност, се получава на единица нестабилност [18]:

$$Balance\ Score_{m,t} = \frac{(T-CPSimprovement\ \%_{m,t}/100)}{CV_{m,t}} \quad \textbf{(3.34)}$$

По-високи стойности на Balance Score означават конфигурации с по-голямо коригирано подобрение при по-ниска нестабилност. Конфигурации с голямо подобрение, но висока нестабилност, получават по-нисък Balance Score от конфигурации с умерено подобрение и ниска нестабилност. Така се предпочитат по-предвидими конфигурации при настройване на извличането.

3.6 Трета глава: Обобщение

Трета глава дефинира групите модели и метриките за оценяване, използвани в Четвърта глава. Подбрани са седем модела с отворен код в диапазона 7-8B, при изискване да са публично достъпни и възможност за локален деплой върху хардуер от среден клас [65], [66]. Оценяването използва панел от 24 метрики (Таблица 3.2), които обхващат лексикално припокриване, семантично сходство, плавност и предсказуемост, качество на отговора, статистическа корелация и показатели за четимост. Обединяването на набор от метрики е дефинирано чрез CPS (3.27)-(3.30), разширено с T-CPS (3.31)-(3.32) и Balance Score (3.34), както и с процедури за статистическа оценка и отчитане на размера на ефекта (3.33), които подпомагат сравнителната интерпретация. Тези компоненти подпомагат адресирането на D1 чрез композитни показатели, съобразени с прага, и на D3 чрез възпроизводими критерии за избор на модели и процедури за оценяване, приложени в Глава 4.

ЧЕТВЪРТА ГЛАВА: ЕКСПЕРИМЕНТАЛНА ОЦЕНКА И АНАЛИЗ

4.1 Фаза I: Тестване на системата и профилиране на времето за изпълнение

Фаза I проверява работата на платформата от "край-до-край" и описва базовото ѝ поведение, като изпълнява пълния RAG процес за три модела със 7B параметри (Mistral 7B, Llama 2 7B, Orca 2 7B) в две хардуерни среди [17]. Извличането е в режим Normal Mode, с фиксиран брой извличани откъси (top-k, K = 100) и температура на генериране 0,2. Фазата е насочена към проверка на инфраструктурния компонент (b) на Obj.1; тестването, отнасящо се до прага на сходство, започва във Фаза II.

4.1.1 Експериментален дизайн

За домейн земеделие са използвани два основни източника: Регламент (ЕС) 2018/848 за биологичното производство [93] и Наръчникът на ФАО за климатично устойчиво земеделие (climate-smart agriculture) [94]. Документите са обработени и сегментирани според параметрите от Секция 2.2.2 (1024 знака, припокриване 50), векторизирани с Mistral 7B и съхранени в ChromaDB. Тествани са 446 двойки въпрос-отговор. Въпросите са генерирани с Mistral 7B, което може да даде предимство на този модел; това е отбелязано в Секция 5.3.2 и е компенсирано във Фаза IV чрез използване на Claude Opus. Използвани са две хардуерни среди: Apple Mac Mini M1 (macOS, 16 GB RAM, с GPU) и Intel Xeon Ubuntu сървър (128 GB RAM, само с централен процесор). Изпълнени са две тестови процедури: RAG Q&A Score Test (16 метрики) и Timing Performance Test. Резултатите са записани в блокчейна чрез смарт контракти (Секция 2.2.3).

4.1.2-4.1.3 Резултати за време на изпълнение и качество

Mac M1 постига приблизително 2.2 пъти по-висока пропускателна способност от Ubuntu конфигурацията само с централен процесор (Таблица 4.1 в дисертацията), докато стойностите на метриките за качество остават съпоставими между двете среди. Малката разлика във времето за зареждане показва, че основното предимство е свързано с бързината на извеждане, а не с инициализацията на модела. При фиксирана конфигурация на извличане Mistral 7B показва най-силни общи резултати по повечето метрики за лексикално и семантично съответствие, Orca 2 7B достига най-високи стойности по ROUGE precision, а Llama 2 7B отчита най-ниска perplexity. Тези резултати показват, че относителното подреждане на моделите зависи от разглежданото семейство метрики. Резултатите от Фаза I имат описателен характер и служат като базова линия за по-късните анализи, чувствителни към прага на сходство.

4.1.4-4.1.5 Анализ и проверка на системата

Пълната верификация на системата потвърждава коректната работа на целия процес - от зареждането на данните и извличането, през генерирането на 1338 отговора и изчисляването на 16 метрики, до експортирането в електронни таблици и регистрирането в блокчейн. Съпоставката между

двете среди подкрепя извода, че качеството на резултатите зависи основно от модела и конфигурацията на извличане, а не от използваната изчислителна среда.

4.1.6 Обобщение на Фаза I

Фаза I потвърждава, че PaSSER работи коректно и че целият процес минава без грешки от извличане до запис на резултатите. Фаза II въвежда филтриране по праг на сходство.

4.2 Фаза II: Праг на сходство и CPS

Фаза II служи като пилотно изследване, чиято цел е да се проследи как прагът на сходство влияе върху качеството на генерирането при ограничена конфигурация на Score Mode и да се оцени CPS като схема за обобщаване преди по-широките анализи във Фази III-IV [15]. Експериментите използват три LLM с 7B параметъра - Mistral 7B, Llama 2 7B и Orca 2 7B - при K = 100, K-Inc = 2 и фиксирана температура 0,2. Прагът на сходство се променя от 0,50 до 0,80 със стъпка 0,05, което води до 2 121 оценки (3 модела × 7 прага × 101 въпроса), изпълнени на Apple Mac mini M1. Този диапазон е избран, за да обхване прехода от по-разрешително към по-селективно извличане, без да се навлиза в по-строги режими на извличане, разгледани по-късно във Фази III-IV. Фаза II разглежда RQ1 в пилотна среда, като прилага компоненти (a) и (b) на Obj. 1 и допринася за Obj. 4.

4.2.1 Експериментален дизайн

Фаза II използва подмножество от 101 двойки въпрос-отговор от набора от данни за Фаза I, като запазва същите настройки за сегментиране, векторизиране и векторно хранилище. Основната промяна в дизайна е въвеждането на извличане в режим Score Mode, което позволява филтриране на извлечените пасажи по праг. Резултати за отделните въпроси са запазени само за прагове 0,50-0,80 и затова всички анализи във Фаза II са ограничени до този интервал.

4.2.2 Стойности на метриките в CPS

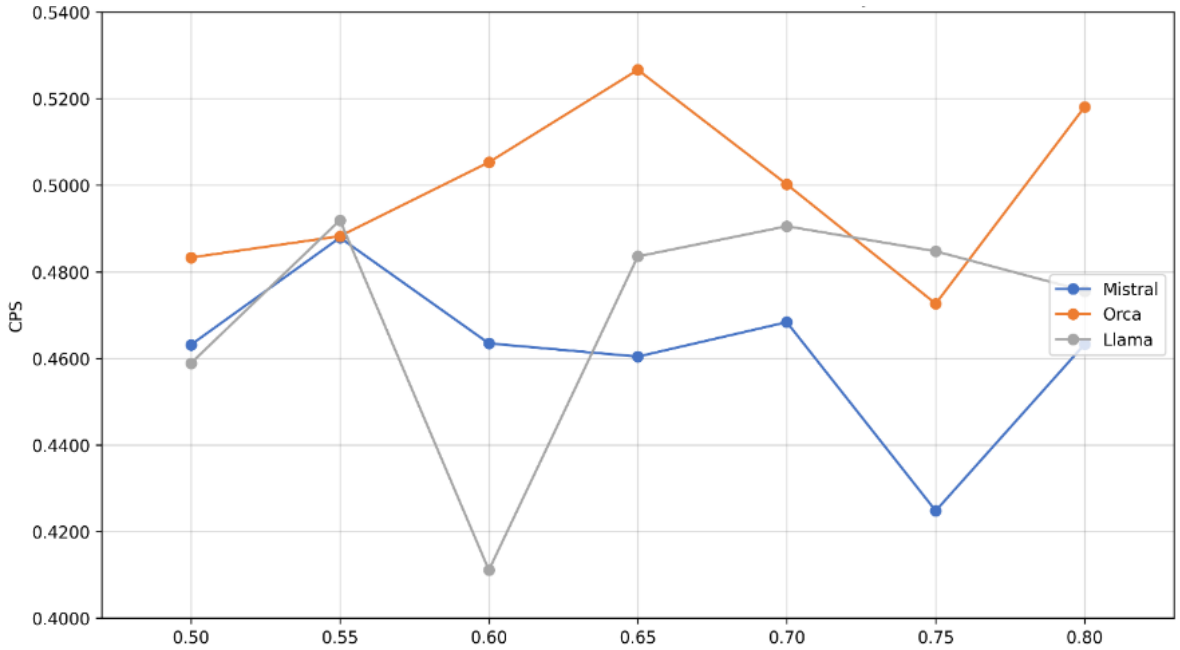
CPS е приложен като претеглено обобщаване на девет метрики, обхващащи четири оценъчни измерения: лексикално припокриване (METEOR, BLEU, ROUGE-1 F, ROUGE-L F), семантично сходство и съгласуваност (Cosine Similarity, Pearson Correlation), плавност и коректност (F1 Score) и езиково моделиране (Laplace Perplexity, Lidstone Perplexity). Метриките за perplexity са с отрицателна полярност и след нормализация се обръщат (Секция 3.5.1). Таблица 4.3 представя набора от метрики и теглата им.

Таблица 4.3. *Метрики и коефициенти в CPS (Фаза II). Източник [15].*

Метрика	Стойност	Обосновка
METEOR	0.15	Цялостна оценка на текстовото качество
ROUGE-1 F-score	0.075	Различни нива на лексикално припокриване
ROUGE-L F-score	0.075	Различни нива на лексикално припокриване
BLEU	0.15	Цялостна оценка на текстовото качество
Laplace Perplexity	0.075	Предсказуемост и точност (по-ниско = по-добре)
Lidstone Perplexity	0.075	Предсказуемост и точност (по-ниско = по-добре)
Cosine Similarity	0.10	Релевантност и съгласуваност при извличане
Pearson Correlation	0.10	Релевантност и съгласуваност при извличане
F1 Score	0.20	Основна метрика за коректност на отговора
ОБЩО	**1.00**	

4.2.3 Резултати

Резултатите от Фаза II са представени описателно, без формално тестване за статистическа значимост, тъй като тази фаза има пилотен характер. В рамките на разглеждания диапазон на прага най-висок CPS е наблюдаван при 0,55 за Mistral 7B и Llama 2 7B и при 0,65 за Orca 2 7B. В целия изследван диапазон Orca 2 7B показва най-стабилен профил на CPS, докато Mistral 7B и Llama 2 7B проявяват по-голяма чувствителност към прага, включително специфични за модела спадове при междинни стойности. Фигура 4.1 илюстрира тези тенденции на CPS при различните прагове и за трите модела.



Фигура 4.1. *CPS стойности за трите модела при праг на сходство 0,50-0,80 (Фаза II, Score Mode). Адаптирана от [15].*

4.2.4-4.2.5 Анализ и обобщение

Пилотното изследване във Фаза II показва, че прагът, при който се наблюдава най-висок CPS, зависи от модела в рамките на изследваната конфигурация. По-нисък или умерен праг може да остане ефективен за модели, които понасят по-успешно по-слабо свързан контекст, докато модели, по-чувствителни към шум при извличането, могат да се повлияват по-добре от по-строго филтриране. Тези тълкувания имат обяснителен, а не причинно-следствен характер, тъй като Фаза II не изолира механизмите, които свързват характеристиките на модела с поведението му спрямо прага. Изводите са ограничени и от условията на пилотното изследване: един домейн, подмножество от 101 въпроса, фиксирани настройки за предварителна обработка и запазени артефакти само за прагове 0,50-0,80. Въпреки тези ограничения Фаза II предоставя начални доказателства по RQ1 и полага основата за по-широките анализи на прага във Фази III-IV.

4.3 Фаза III: Прагове на сходство според модела

Фаза III изследва как се променя представянето на езиковите модели при промяна на прага на сходство в RAG, с акцент върху чувствителността, зависеща от модела, при все по-селективно извличане. Групата модели е обновена спрямо Фаза II: Mistral 7B е запазен за приемственост в актуализирана версия 0.3, а Llama 2 7B и Orca 2 7B са заменени с три модела от клас с 8B параметра - DeepSeek R1 8B, Llama 3.1 8B и Granite 3.2 8B. Оценени са прагове на сходство от 0,50 до 0,95 със стъпка 0,05. Качеството на генерирането е обобщено чрез CPS и T-CPS, като за всяко изпълнение е изчислен пълният набор от 24 метрики. Базовата конфигурация е Normal Mode (top-k извличане без филтриране по праг), а тестовете с праг използват Score Mode. Фаза III разглежда RQ1 и RQ2, като прилага компоненти (a) и (b) на Obj. 1 и допринася за Obj. 4.

4.3.1 Експериментален дизайн

Използвано е подмножество от 369 двойки въпрос-отговор от набора от данни за Фаза I, като настройките за сегментиране, векторизиране и векторно хранилище са запазени без промяна. Експериментите са проведени в три хардуерни среди - M1 Mac Mini, M2 Mac Mini и сървър само с централен процесор. Размерът на контекстния буфер се различава по модел (2 048-10 000 токена) поради ограничения в паметта; отражението на това е разгледано в Секция 5.3.2. За да се ограничат страничните ефекти, свързани с хардуера, статистическите сравнения са проведени в рамките на всеки модел спрямо собствената му базова конфигурация. Общо са изпълнени 16 236 оценки (4 модела × 11 конфигурации × 369 двойки въпрос-отговор).

4.3.2 Стойности на метриките в CPS

Във Фаза III се прилагат CPS и T-CPS с обновен панел от метрики, който разширява семантичното и езиковото покритие: 30% лексикално припокриване (METEOR, ROUGE), 25% семантично сходство (BERTScore.f1, B-RT.average), 25% плавност и точност (F1 Score, B-RT.fluency) и 20% езиково моделиране

(Laplace/Lidstone Perplexity). Таблица 4.6 описва промените в панела от Фаза II към Фази III-IV.

Таблица 4.6 *Развитие на панела от метрики за CPS между експерименталните фази.*

Категория	Метрика	Фаза II	Фаза III-IV	Промяна	Обосновка
Лексикално припокриване	METEOR	0,150	0,150	—	Запазена основна метрика
	BLEU	0,150	—	Премахната	Припокриване с METEOR
	ROUGE-1.f	0,075	—	Заменена	Униграм по-малко информативен от биграм
	ROUGE-2.f	—	0,075	Добавена	Биграмно припокриване
	ROUGE-L.f	0,075	0,075	—	Запазена
	Междинен сбор	**0,450**	**0,300**	**−0,150**	
Семантично сходство	Cosine Similarity	0,100	—	Премахната	Заменена от контекст. вграждания
	Pearson Correlation	0,100	—	Премахната	По-малко интерпретируема
	BERTScore.f1	—	0,125	Добавена	Контекстуално токенно сходство
	B-RT.average	—	0,125	Добавена	Многоаспектен показател за четимост
	Междинен сбор	**0,200**	**0,250**	**+0,050**	
Плавност и точност	F1 Score	0,200	0,150	−0,050	Тегло преразпределено
	B-RT.fluency	—	0,100	Добавена	Оценка за плавност
	Междинен сбор	**0,200**	**0,250**	**+0,050**	
Езиково моделир.	Laplace Perplexity	0,075	0,100	+0,025	Повишен акцент
	Lidstone Perplexity	0,075	0,100	+0,025	Повишен акцент
	Междинен сбор	**0,150**	**0,200**	**+0,050**	
ОБЩО		**1,000**	**1,000**		

4.3.3 Обобщение на резултатите по CPS

Таблица 4.7 обобщава конфигурациите с най-висок CPS подобрение за всеки модел. Най-високите CPS подобрения са модел-зависими: Mistral 7B v0.3 постига най-голямо подобрение (+4,58% при праг 0,95), следван от Llama 3.1 8B (+1,58% при 0,90), Granite 3.2 8B (+1,25% при 0,95) и DeepSeek R1 8B (+1,01% при 0,90).

Таблица 4.7 *Конфигурации с най-високо CPS подобрение по модели (Топ 3, земеделие).*

Модел	Ранг	Праг	Среден CPS	Подобр. %
Mistral 7B v0.3	1	0,95	0,5454	**4,58**
	2	0,90	0,5338	2,37
	3	0,70	0,5325	2,11
Granite 3.2 8B	1	0,95	0,5182	**1,25**
	2	0,70	0,5179	1,20
	3	0,80	0,5178	1,17
Llama 3.1 8B	1	0,90	0,5080	**1,58**

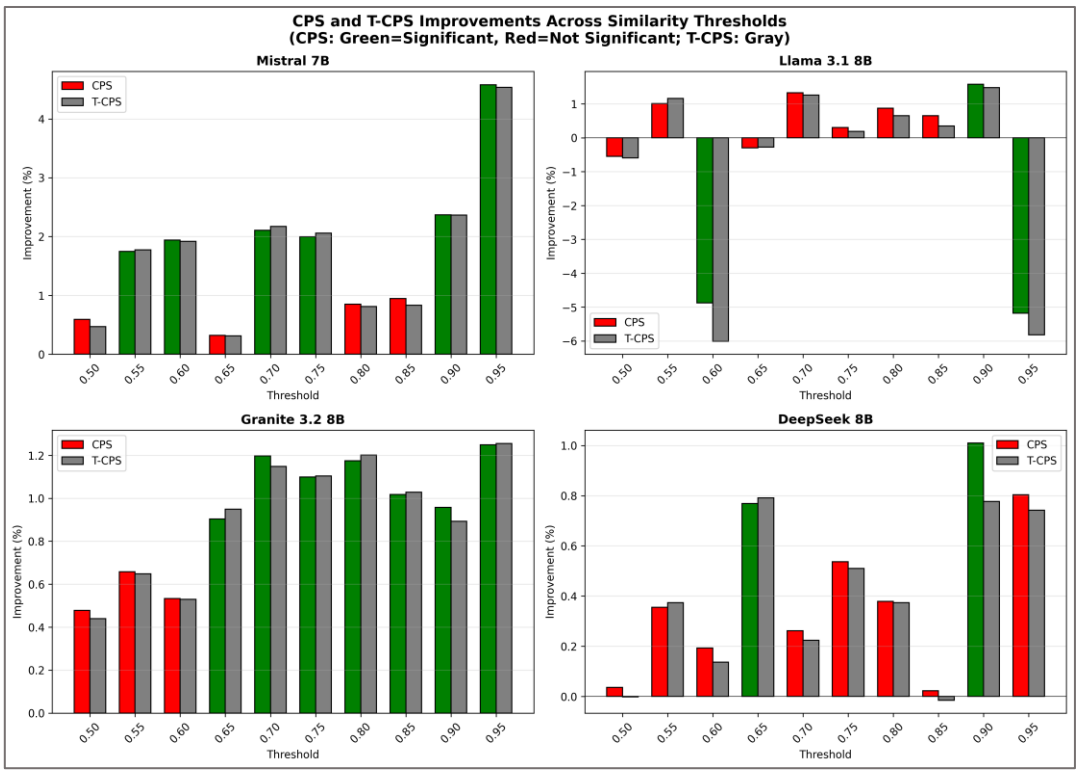
	2	0,70	0,5065	1,33
	3	0,55	0,5052	1,01
DeepSeek R1 8B	1	0,90	0,4559	**1,01**
	2	0,95	0,4550	0,80
	3	0,65	0,4548	0,77

4.3.4 Представяне по T-CPS и стабилност

Таблица 4.8 представя конфигурациите с най-висок T-CPS за всеки модел. Класирането по T-CPS до голяма степен повтаря класирането по CPS: Mistral 7B v0.3 достига максимум при 0,95 (+4,54%), Granite 3.2 8B при 0,95 (+1,25%), Llama 3.1 8B при 0,90 (+1,48%). DeepSeek R1 8B се отличава, с максимум при праг 0,65 (+0,79%) вместо оптималния по CPS праг 0,90, което отразява компромис между стабилност и производителност. DeepSeek R1 8B показва по-ниска нестабилност (CV = 0,085–0,108) в сравнение с останалите модели (CV = 0,122–0,148). Фигура 4.2 сравнява CPS и T-CPS подобренията по прагове за всеки модел.

Таблица 4.8 *Конфигурации с най-висок T-CPS по модел (Топ 3, земеделие).*

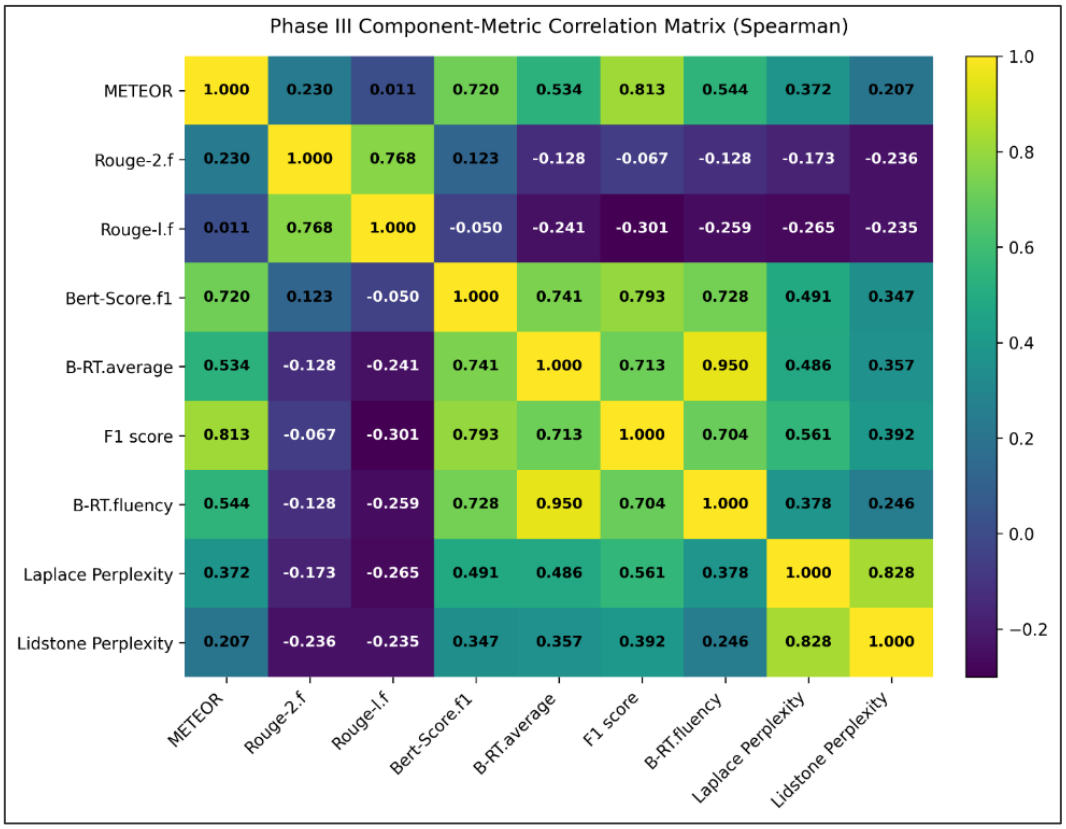
Модел	Ранг	Праг	T-CPS	T-CPS %	CV	Тълкуване
Mistral 7B v0.3	1	0,95	0,5916	4,54	0,134	Голямо подобрение
	2	0,90	0,5793	2,36	0,131	Умерено подобрение
	3	0,70	0,5783	2,17	0,128	Умерено подобрение
Granite 3.2 8B	1	0,95	0,5628	1,25	0,124	Малко подобрение
	2	0,80	0,5625	1,20	0,122	Малко подобрение
	3	0,70	0,5622	1,15	0,129	Малко подобрение
Llama 3.1 8B	1	0,90	0,5501	1,48	0,148	Малко подобрение
	2	0,70	0,5490	1,26	0,142	Малко подобрение
	3	0,55	0,5484	1,16	0,129	Малко подобрение
DeepSeek R1 8B	1	0,65	0,4961	0,79	0,085	Малко подобрение
	2	0,90	0,4960	0,78	0,108	Малко подобрение
	3	0,95	0,4958	0,74	0,093	Малко подобрение



Фигура 4.2 Фаза III (земеделие, N = 369): CPS и T-CPS подобрения по прагове за всеки модел.

4.3.5 Корелационен анализ

Корелационният анализ оценява съвпадението между метриките и дали T-CPS дава информация, допълнителна към средния CPS. Фигура 4.3 показва корелационната матрица (Spearman) между метриките от Фаза III, обединени по модели и прагове 0,50–0,95 плюс базовата линия. Таблица 4.9 отчита зависимости между T-CPS и CPS (ρ ≈ 0,999) и между T-CPS и CV (ρ ≈ 0,392), което потвърждава, че T-CPS следва отблизо средната производителност, като внася вторично предпочитание за стабилност.



Фигура 4.3 *Корелационна матрица (Spearman) на метриките от Фаза III.*

Таблица 4.9 *Зависимости между T-CPS, CPS и CV (Spearman; N = 44 конфигурации).*

Зависимост	Spearman ρ
ρ(T-CPS, CPS)	0,999
ρ(T-CPS, CV)	0,392

4.3.6 Balance Score

Balance Score класира конфигурациите по подобрение, коригирано за стабилност, на единица нестабилност (Секция 3.5.4). Таблица 4.10 показва 10-те най-добри статистически значими положителни конфигурации. Mistral 7B v0.3 доминира класирането поради по-големи T-CPS подобрения при умерен CV, докато DeepSeek R1 8B постига конкурентни Balance Score стойности въпреки по-малките подобрения благодарение на нисък CV.

Таблица 4.10 *Класиране по Balance Score (Топ 10 значими положителни конфигурации).*

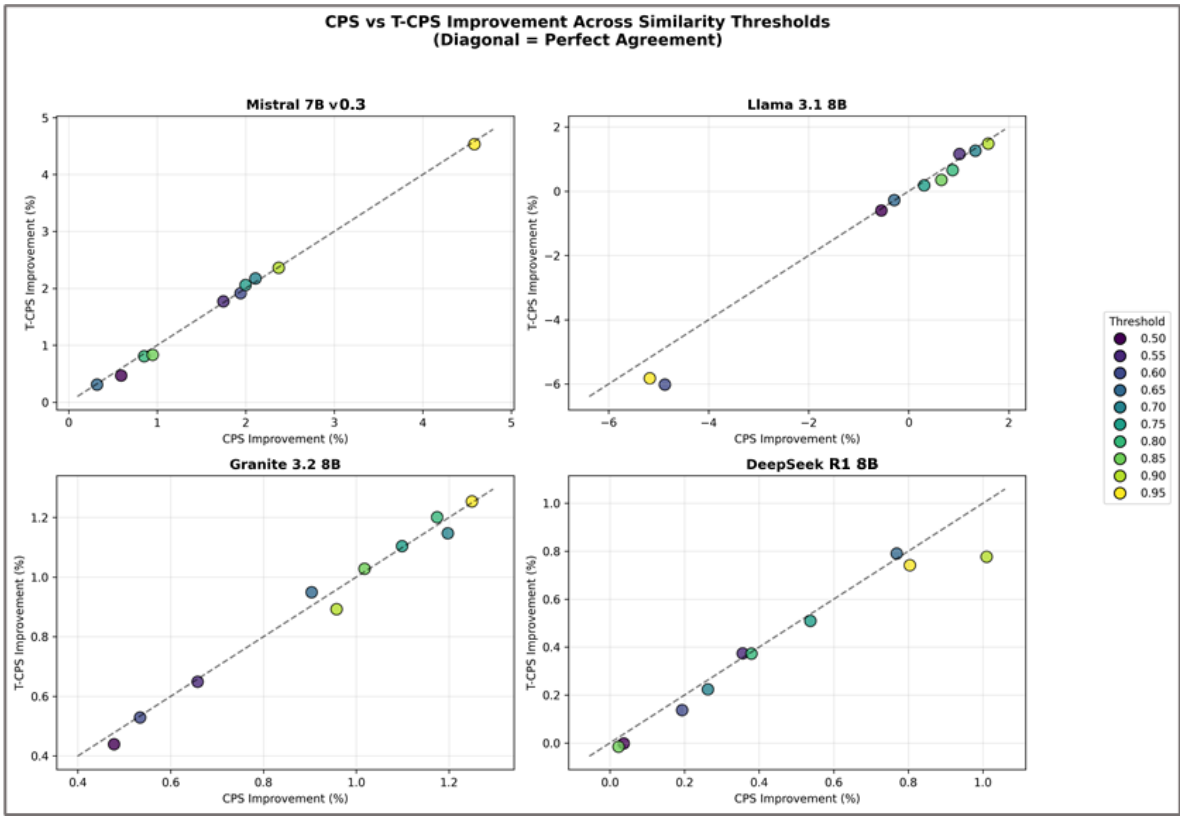
#	Модел	Праг	T-CPS %	CV	Balance Score	Знач.
1	Mistral 7B v0.3	0,95	+4,54	0,134	0,339	***
2	Mistral 7B v0.3	0,90	+2,36	0,131	0,180	**
3	Mistral 7B v0.3	0,70	+2,17	0,128	0,170	*
4	Mistral 7B v0.3	0,75	+2,06	0,128	0,161	*
5	Mistral 7B v0.3	0,60	+1,92	0,135	0,142	*
6	Mistral 7B v0.3	0,55	+1,77	0,132	0,134	*
7	Granite 3.2 8B	0,95	+1,25	0,124	0,101	**
8	Llama 3.1 8B	0,90	+1,48	0,148	0,100	*
9	Granite 3.2 8B	0,80	+1,20	0,122	0,098	**
10	DeepSeek R1 8B	0,65	+0,79	0,085	0,093	**

Таблица 4.11 обобщава съвпадението между праговете по най-висок CPS, най-висок T-CPS и най-висок Balance Score. За Mistral 7B v0.3, Granite 3.2 8B и Llama 3.1 8B трите критерия сочат един и същ праг, а при DeepSeek R1 8B оптималните стойности се разминават (CPS при 0,90 срещу CV при 0,65).

Таблица 4.11. *Съвпадение на праговете между критериите за избор (по модели).*

Модел	Най-добър праг по CPS	Най-добър праг по T-CPS	Най-добър праг по Balance Score	Alignment
Mistral 7B v0.3	0,95	0,95	0,95	Пълно съвпадение
Granite 3.2 8B	0,95	0,95	0,95	Пълно съвпадение
Llama 3.1 8B	0,9	0,9	0,9	Пълно съвпадение
DeepSeek R1 8B	0,9	0,65	0,65	Разминаване

Фигура 4.4 показва съвпадението между CPS и T-CPS по прагове за всеки модел.



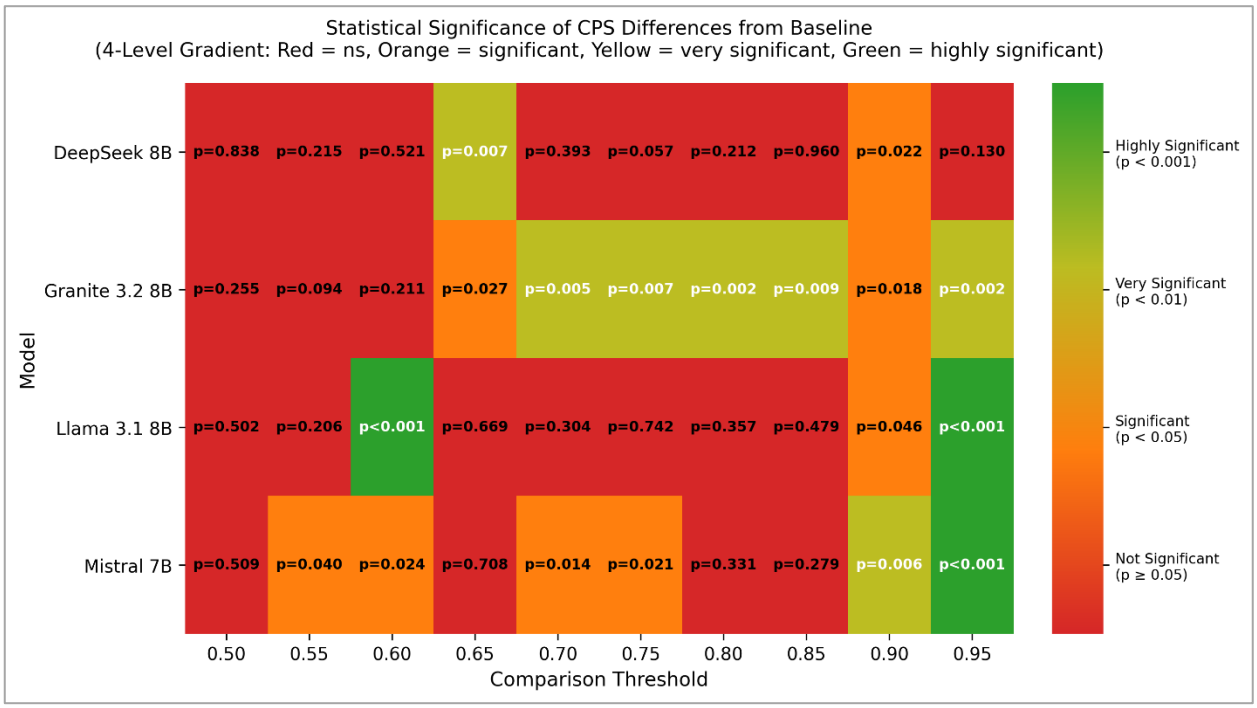
Фигура 4.4*. Фаза III (земеделие, N = 369): съвпадение CPS-T-CPS по прагове за всеки модел (диагонал = пълно съвпадение).*

4.3.7 Статистическа значимост

Направени са t-тестове по двойки (paired), като за всеки праг CPS на ниво въпрос е сравнен с базовата линия (без корекция). Mistral 7B v0.3 показва значими положителни подобрения при 6 от 10 прага, Granite 3.2 8B при 7 от 10, Llama 3.1 8B — едно значимо подобрение и два значими спада, а DeepSeek R1 8B — подобрения при два прага. Таблица 4.12 обобщава разпределението на значимостта; Фигура 4.5 представя термокарта на значимостта по прагове.

Таблица 4.12 *Разпределение на значимостта по модел (домейн земеделие).*

Модел	Знач. положит.	Знач. отрицат.	Незначим.
Mistral 7B v0.3	6	0	4
Granite 3.2 8B	7	0	3
Llama 3.1 8B	1	2	7
DeepSeek R1 8B	2	0	8



Фигура 4.5 *Термокарта на статистическата значимост на CPS разликите спрямо базовата линия (Фаза III).*

4.3.8 Чувствителност към прага по модели

Mistral 7B v0.3 показва най-благоприятния профил, със значими подобрения в обхвата 0,55–0,95 и максимум при 0,95. Granite 3.2 8B демонстрира широк ефективен обхват (0,65–0,95) при сравнително слаби колебания. Llama 3.1 8B има тесен ефективен обхват с висока чувствителност към избора на праг. DeepSeek R1 8B показва ограничени подобрения и разминаване между оптималните прагове по средна производителност и по показателите, отчитащи стабилността.

4.3.9 Обобщение на Фаза III

Таблица 4.13 обобщава най-добрите конфигурации по модел с маркери за значимост. Оптималните прагове варират: Mistral 7B v0.3 и Granite 3.2 8B при 0,95, Llama 3.1 8B при 0,90, DeepSeek R1 8B с разминаващи се оптимуми (0,90 срещу 0,65). Резултатите са ограничени от тестваната конфигурация (четири модела, N = 369, домейн земеделие, разнороден хардуер). Данните осигуряват доказателства за RQ1 и RQ2 в домейна на земеделието; преносимостта между домейни се проверява във Фаза IV.

Таблица 4.13 *Обобщение на най-добрите конфигурации (домейн земеделие).*

Модел	Праг	CPS %	T-CPS %	CV	Balance	Знач.
Mistral 7B v0.3	0,95	4,58	4,54	0,134	0,339	***
Granite 3.2 8B	0,95	1,25	1,25	0,124	0,101	**
Llama 3.1 8B	0,90	1,58	1,48	0,148	0,100	*
DeepSeek R1 8B	0,65/0,90	0,77/1,01	0,79/0,78	0,085/0,108	0,093/0,072	**/ *

4.4 Фаза IV: Проверка на преносимостта между домейни (Биоразнообразие)

Фаза IV разширява анализа на прага на сходство от Фаза III към домейна на биоразнообразието, за да се изследва преносимостта между домейни. Групата модели, диапазонът на прага (0,50-0,95 със стъпка 0,05), начинът на извличане, форматът на промпта, настройките за сегментиране, моделът за векторизация, температурата и рамката за оценка остават непроменени. Основните промени са свързани с корпуса от знания, тестовия набор по биоразнообразие (N = 426 двойки въпрос-отговор) и използването на една-единствена среда за изпълнение (M1 Mac Mini). За всяко изпълнение е изчислен пълният набор от 24 метрики. Целта е да се установи дали ефектите на прага и моделно-зависимите профили на чувствителност се запазват, когато характеристиките на корпуса са различни, включително по отношение на речника и разпределенията на сходство във векторните представяния. Фаза IV разглежда RQ1, RQ2 и RQ3, като повтаря процедурата от Фаза III във втора тематична област и сравнява получените зависимости между прага и отговора на моделите.

4.4.1 Експериментален дизайн

Всички експерименти са проведени в една среда за изпълнение - M1 Mac Mini - при фиксиран контекстен буфер от 16 000 токена за всички модели и прагове. Тази стандартизация премахва вариацията в контекстния буфер като смущаващ фактор и засилва основанията наблюдаваните различия да се тълкуват като домейнно обусловени, а не като свързани с хардуера. Корпусът по биоразнообразие включва 426 двойки въпрос-отговор, изведени от авторитетни източници, сред които Конвенцията за биологичното разнообразие [95] и Стратегията на Европейския съюз за биологичното разнообразие [96]. Референтните отговори са извлечени от изходните документи, а въпросите са генерирани с Claude Opus по същата обща процедура, използвана във Фаза I. Общо са проведени 18 744 оценки (4 модела × 11 конфигурации × 426 двойки въпрос-отговор).

4.4.2 Стойности на метриките в CPS

Без промяна е запазена същата схема за претегляне с 9 метрики, използвана във Фаза III (Таблица 4.6). Схемата на теглата следва рамката за оценяване с четири измерения: лексикално припокриване (30%), семантично сходство (25%), плавност и точност (25%) и езиково моделиране (20%).

4.4.3 Обобщение на резултатите по CPS

Представянето по CPS в интервала 0,50-0,95 е оценено за всички четири модела в домейна на биоразнообразието. Таблица 4.14 представя трите конфигурации с най-голямо подобрение на CPS за всеки модел. В сравнение с домейна на земеделието резултатите при домейн биоразнообразие показват по-големи повишения на CPS и по-ниски прагове на максимум при всички модели. Най-високите подобрения на CPS са наблюдавани при Mistral 7B v0.3 при 0,80 (+13,32%), DeepSeek R1 8B при 0,55 (+8,45%), Granite 3.2 8B при 0,80 (+6,95%) и Llama 3.1 8B при 0,85 (+2,06%).

Таблица 4.14 *Конфигурации с най-високо CPS подобрение по модел (Топ 3, биоразнообразие).*

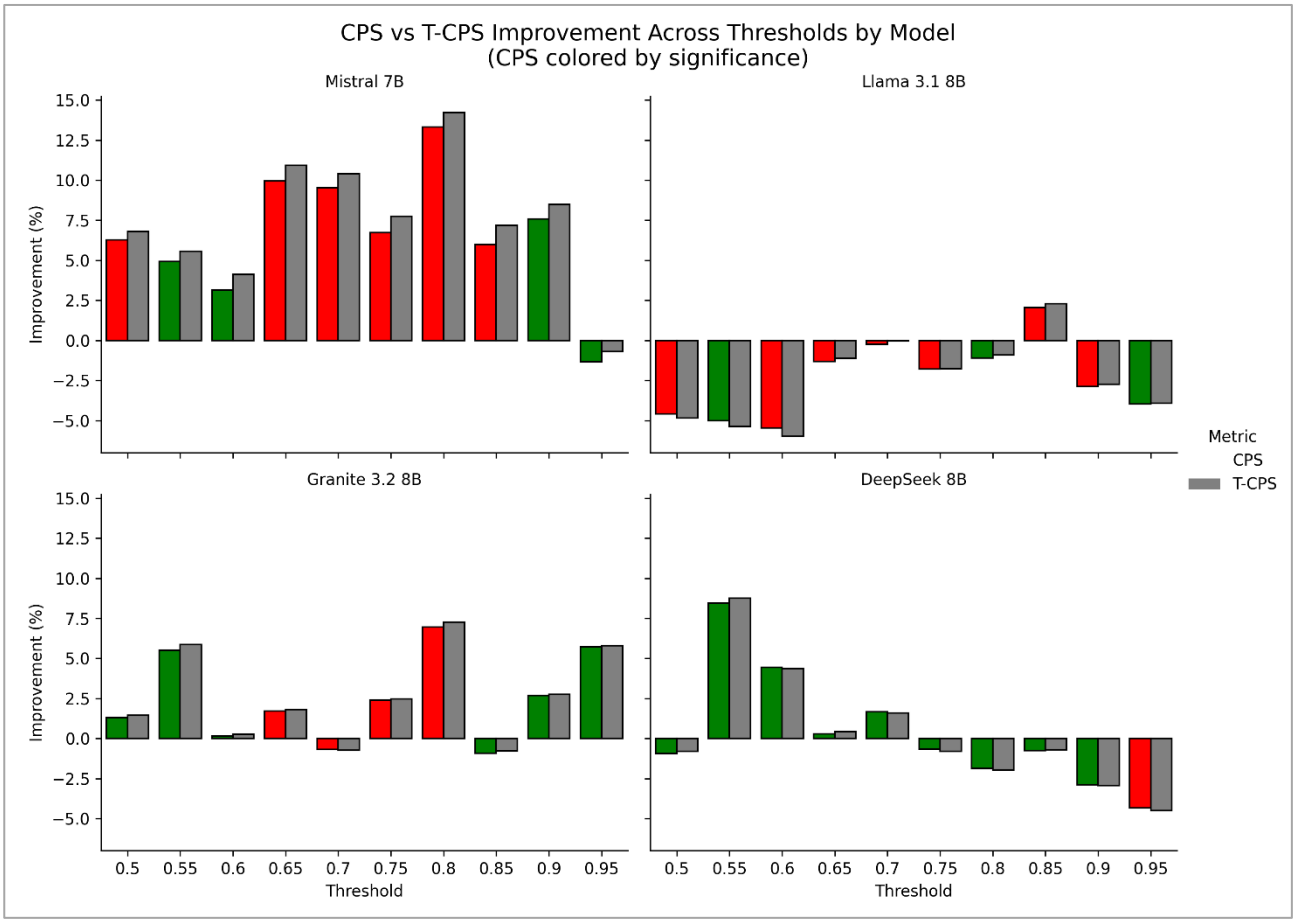
Модел	Ранг	Праг	Среден CPS	Подобр. %
Mistral 7B v0.3	1	0,80	0,4911	**13,32**
	2	0,65	0,4764	9,94
	3	0,70	0,4747	9,53
Granite 3.2 8B	1	0,80	0,4473	**6,95**
	2	0,95	0,4422	5,73
	3	0,55	0,4413	5,51
Llama 3.1 8B	1	0,85	0,4713	**2,06**
	2	0,70	0,4606	−0,25
	3	0,80	0,4567	−1,11
DeepSeek R1 8B	1	0,55	0,5094	**8,45**
	2	0,60	0,4906	4,45
	3	0,70	0,4775	1,66

4.4.4 Представяне по T-CPS и стабилност

Таблица 4.15 обобщава конфигурациите с най-висок T-CPS за всеки модел, като отчита и изменението чрез коефициента на вариация (CV). Класирането с отчитане на стабилността остава зависимо от модела, но в голяма степен съвпада с резултатите по CPS: Mistral 7B v0.3 и Granite 3.2 8B достигат максимум при 0,80, Llama 3.1 8B - при 0,85, а DeepSeek R1 8B - при 0,55. Сред водещите конфигурации DeepSeek R1 8B показва най-слаби колебания(CV = 0,129-0,158), докато останалите модели остават в по-висок диапазон (CV = 0,233-0,254), което показва по-стабилно качество на отговорите му при различните заявки. Тази зависимост се вижда и на Фигура 4.6, която сравнява подобренията по CPS и T-CPS за всеки модел по прагове и показва къде отчитането на стабилността променя подреждането спрямо подобрението само по CPS.

Таблица 4.15 *Конфигурации с най-високо T-CPS подобрение по модел (Топ 3, биоразнообразие).*

Model	Rank	Threshold	T-CPS	T-CPS Impr. %	CV	Interpretation
Mistral 7B v0.3	1	0.8	0.5254	14.23	0.242	Large improvement
	2	0.65	0.5102	10.93	0.233	Large improvement
	3	0.7	0.5078	10.41	0.24	Large improvement
Granite 3.2 8B	1	0.8	0.4785	7.25	0.239	Moderate improvement
	2	0.55	0.4723	5.87	0.235	Moderate improvement
	3	0.95	0.472	5.8	0.254	Moderate improvement
Llama 3.1 8B	1	0.85	0.5042	2.29	0.24	Small improvement
	2	0.7	0.4928	−0.03	0.24	Minimal change
	3	0.8	0.4884	−0.92	0.241	Minimal change
DeepSeek R1 8B	1	0.55	0.5529	8.75	0.129	Large improvement
	2	0.6	0.5306	4.38	0.158	Moderate improvement
	3	0.7	0.5165	1.59	0.157	Small improvement

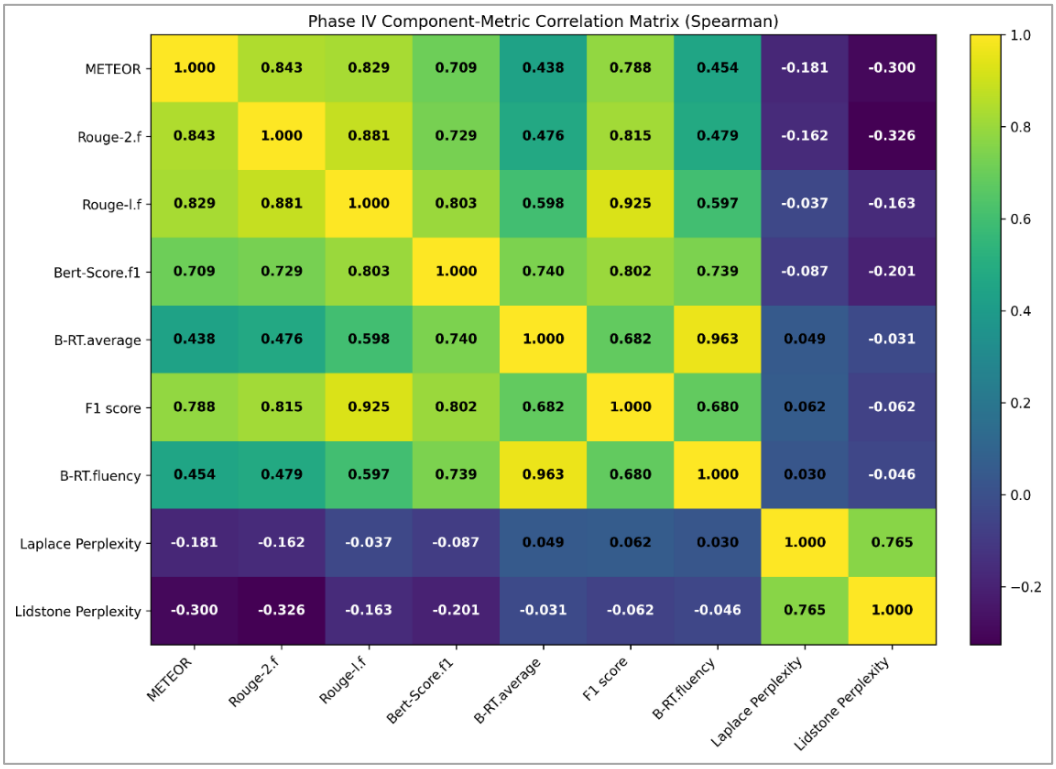


Фигура 4.6 Фаза IV (биоразнообразие, N = 426): CPS и T-CPS подобрения по прагове за всеки модел.

4.4.5 Корелационен анализ

Корелационният анализ е използван, за да се оцени припокриването между метриките и да се установи дали T-CPS добавя информация отвъд средния CPS. Фигура 4.7 представя корелационната матрица (Spearman) за компонентните метрики от Фаза IV, обединени за праговете 0,50-0,95 и базовата конфигурация. Таблица 4.16 показва, че T-CPS остава силно свързан с CPS (ρ ≈ 0,992), като същевременно показва и съществена обратна зависимост с коефициента на вариация (CV) (ρ ≈ -0,724). Това показва, че във Фаза IV T-CPS запазва подреждането по средно представяне, като едновременно с това изразява по-силен сигнал за стабилност, отколкото във Фаза III. Корелационната структура допълнително показва по-плътно групиране на метриките за лексикално припокриване и домейнно

обусловено изместване в зависимостите, свързани с perplexity, което показва, че взаимозависимостите между метриките са чувствителни към характеристиките на корпуса.



Фигура 4.7. *Корелационна матрица (Spearman) на метриките от Фаза IV.*

Table 4.16 *Зависимости между T-CPS, CPS и CV (Spearman; N = 44 конфигурации).*

Зависимост	Spearman ρ
ρ(T-CPS, CPS)	0,992
ρ(T-CPS, CV)	-0,724

4.4.6 Balance Score

Таблица 4.17 представя 10-те най-добри значими конфигурации. DeepSeek R1 8B при праг 0,55 постига най-висок Balance Score (0,678) поради силно подобрение (+8,75%) и нисък CV (0,129), докато Mistral 7B v0.3 заема множество позиции поради стабилно високи подобрения. Таблица 4.18 показва пълно съвпадение на оптималния праг по CPS, T-CPS и Balance Score за всички четири модела в домейна биоразнообразие, което опростява избора в сравнение с Фаза III.

Table 4.17 *Класиране по Balance Score (Топ 10 значими положителни конфигурации).*

#	Модел	Праг	T-CPS %	CV	Balance Score	Знач.
1	DeepSeek R1 8B	0.55	+8.75	0.129	0.678	***
2	Mistral 7B v0.3	0.8	+14.23	0.242	0.588	***
3	Mistral 7B v0.3	0.65	+10.93	0.233	0.469	***
4	Mistral 7B v0.3	0.7	+10.41	0.24	0.434	***
5	Mistral 7B v0.3	0.9	+8.49	0.236	0.36	***
6	Mistral 7B v0.3	0.75	+7.74	0.228	0.34	***
7	Mistral 7B v0.3	0.85	+7.17	0.217	0.33	***
8	Granite 3.2 8B	0.8	+7.25	0.239	0.303	***
9	DeepSeek R1 8B	0.6	+4.38	0.158	0.277	***
10	Mistral 7B v0.3	0.5	+6.8	0.26	0.262	***

Таблица 4.18 показва пълно съвпадение между CPS, T-CPS и Balance Score за всички четири

модела в домейна на биоразнообразието, което прави избора на праг по-еднозначен, отколкото във Фаза III.

Таблица 4.18 *Съвпадение на оптималния праг между критериите за подбор по модел.*

Модел	Най-добър CPS праг	Най-добър T-CPS праг	Най-добър Balance Score праг	Съвпадение
Mistral 7B v0.3	0,80	0,80	0,80	Пълно
Granite 3.2 8B	0,80	0,80	0,80	Пълно
Llama 3.1 8B	0,85	0,85	0,85	Пълно
DeepSeek R1 8B	0,55	0,55	0,55	Пълно

Фигура 4.8 сравнява подобрението по CPS с подобрението по T-CPS за всеки праг и модел. Точките по диагонала показват пълно съвпадение, а отклоненията от него показват къде отчетената нестабилност променя оценката спрямо средната стойност.

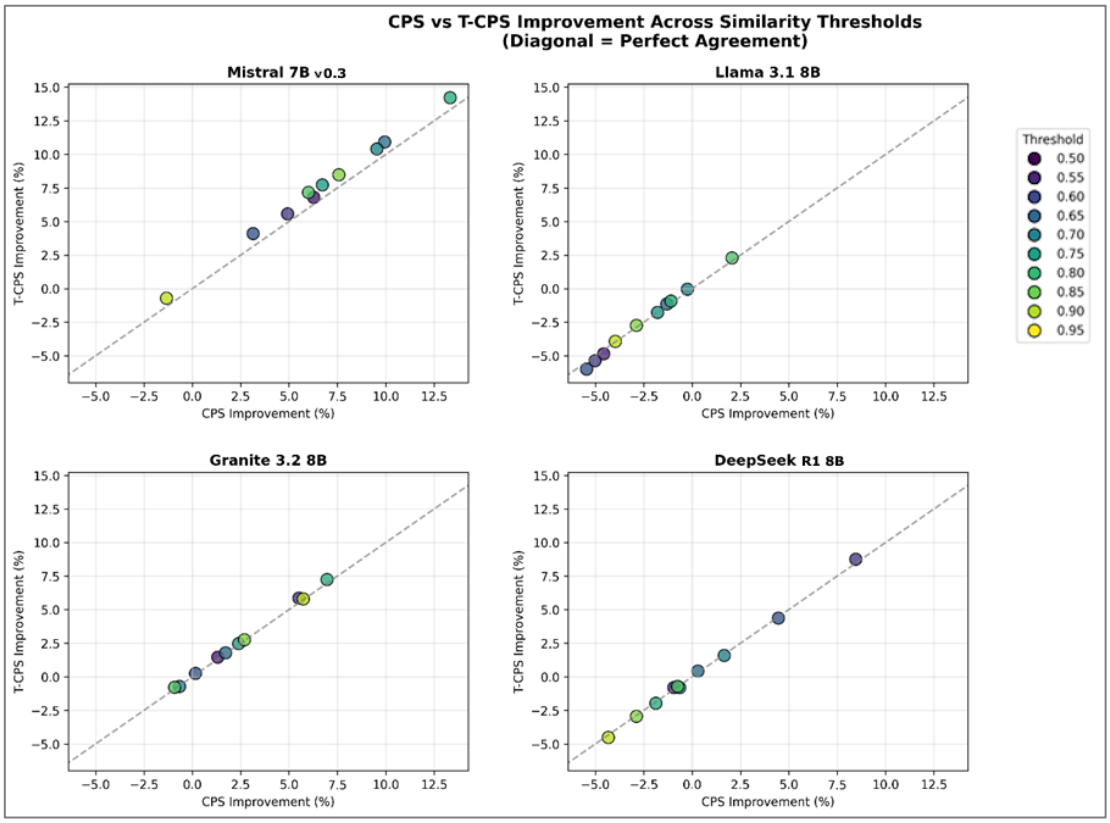


Figure 4.8 *Фаза IV (биоразнообразие, N = 426): съвпадение CPS-T-CPS по прагове за всеки модел (диагонал = пълно съвпадение).*

4.4.7 Статистическа значимост

Проведени са двустранни сдвоени t-тестове, при които стойностите на CPS за отделните въпроси при всеки праг на сходство са сравнени с базовата конфигурация без корекция за множествени сравнения. Таблица 4.19 обобщава разпределението на статистическата значимост по модел, а Фигура 4.9 представя съответната термокарта. Mistral 7B v0.3 показва статистически значими положителни подобрения при 9 от 10 прага, което очертава най-широкия ефективен диапазон в този домейн. Granite 3.2 8B показва статистически значимо подобрение при 3 прага, DeepSeek R1 8B - при 2 прага, заедно с 2 статистически значими отрицателни ефекта при по-строги настройки, а Llama 3.1 8B показва само 1 статистически значим положителен резултат наред с множество статистически значими понижения.

Table 4.19 *Разпределение на значимостта по модел (домейн биоразнообразие).*

Модел	Знач. положит.	Знач. отрицат.	Незначим.
DeepSeek R1 8B	2	2	6
Granite 3.2 8B	3	0	7

Llama 3.1 8B	1	5	4
Mistral 7B v0.3	9	0	1

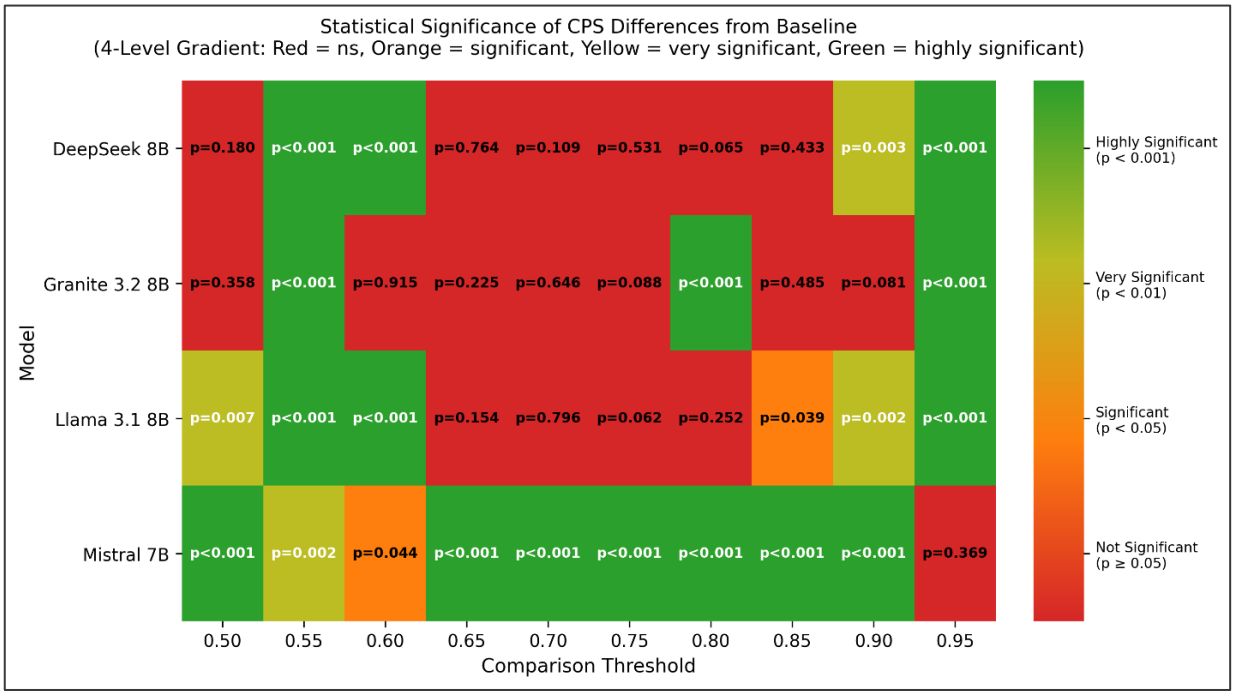


Figure 4.9 *Термокарта на статистическата значимост на CPS разликите спрямо базовата линия (Фаза IV).*

4.4.8 Чувствителност към прага по модели

Модел-зависимите профили на чувствителност се запазват и при смяна на домейна, но ефективният обхват на прага и степента на подобрение се променят. В домейна биоразнообразие Mistral 7B v0.3 показва най-широк положителен отговор, с максимум при 0,80 (+13,32% CPS; +14,23% T-CPS) и подобрения при повечето от изследваните прагове. Granite 3.2 8B показва по-умерен отговор, със значими подобрения при по-малък брой прагове и най-високо представяне също при 0,80 (+6,95% CPS; +7,25% T-CPS). Llama 3.1 8B остава силно чувствителен към избора на праг, с единствен значим положителен праг при 0,85 и няколко прага, свързани със значимо влошаване. DeepSeek R1 8B показва ефективен обхват, съсредоточен при по-ниски прагове, особено 0,55-0,60, със силни подобрения при 0,55, но с отрицателни ефекти при по-строго филтриране. Той постига и най-висок Balance Score в тази фаза.

4.4.9 Сравнение между тематичните области

Сравнението между Фаза IV и Фаза III показва зависими от тематичната област промени в оптималните стойности на прага и в големината на отчетените подобрения. И при четирите модела праговете, при които се постигат най-добри резултати по CPS, са по-ниски в областта на биоразнообразието, отколкото в земеделието: при Mistral 7B v0.3 и Granite 3.2 8B се наблюдава изместване от 0,95 на 0,80, при Llama 3.1 8B - от 0,90 на 0,85, а при DeepSeek R1 8B - от 0,90 на 0,55. Големината на подобренията също нараства при биоразнообразието, което показва по-силна чувствителност към прага в този корпус. Изборът на праг става и по-последователен: и четирите модела показват пълно съвпадение между CPS, T-CPS и Balance Score, докато при земеделието DeepSeek R1 8B изисква компромис между средното представяне и показателите, отчитащи стабилността. В обобщение, резултатите показват, че праговете на сходство с най-добри резултати зависят от тематичната област и трябва да се настройват спрямо целевия корпус, вместо да се пренасят без промяна между различни области.

Таблица 4.20 *Сравнение между домейни: земеделие (Фаза III) срещу биоразнообразие (Фаза IV).*

Модел	Земед. праг	Земед. CPS %	Биоразн. праг	Биоразн. CPS %	Изместване
Mistral 7B v0.3	0,95	+4,58%	0,80	+13,32%	−0,15

Granite 3.2 8B	0,95	+1,25%	0,80	+6,95%	−0,15
Llama 3.1 8B	0,90	+1,58%	0,85	+2,06%	−0,05
DeepSeek R1 8B	0,90	+1,01%	0,55	+8,45%	−0,35

4.4.10 Обобщение на Фаза IV

Фаза IV оценява чувствителността към прага при четири модела с отворен код, използвайки 426 двойки въпрос-отговор от домейна биоразнообразие при стандартизиран хардуер. Таблица 4.21 обобщава конфигурациите с най-добри резултати по модел. Най-добрите прагове се изместват надолу спрямо земеделието, а подобренията са по-големи: Mistral 7B v0.3 при 0,80 (CPS +13,32%; T-CPS +14,23%), Granite 3.2 8B при 0,80 (+6,95%; +7,25%), Llama 3.1 8B при 0,85 (+2,06%; +2,29%), DeepSeek R1 8B при 0,55 (+8,45%; +8,75%). Данните осигуряват доказателства за RQ1 и RQ2 в домейна биоразнообразие и за RQ3 чрез сравнение между домейни с Фаза III, подкрепяйки извода, че праговете на сходство изискват калибриране за конкретния корпус.

Таблица 4.21 *Обобщение на най-добрите конфигурации (домейн биоразнообразие).*

Модел	Праг	CPS %	T-CPS %	CV	Balance	Знач.
Mistral 7B v0.3	0,80	13,32	14,23	0,242	0,588	***
Granite 3.2 8B	0,80	6,95	7,25	0,239	0,303	***
Llama 3.1 8B	0,85	2,06	2,29	0,240	0,095	*
DeepSeek R1 8B	0,55	8,45	8,75	0,129	0,678	***

4.5 Четвърта глава: Обобщение

Глава 4 разглежда как прагът на сходство влияе върху качеството на генериране в RAG при седем езикови модела с отворен код и в две тематични области, като използва платформата PaSSER, представена във Втора глава, и рамката за оценка, дефинирана в Трета глава. Фаза I установява базовото поведение на системата при фиксирано извличане top-k и сравнява резултатите за време на изпълнение и качество при три 7B модела в две хардуерни среди. Фаза II въвежда извличане, базирано на праг на сходство. Пилотно изследване в диапазона от 0,50 до 0,80 и показва, че CPS се изменя според модела. Фаза III разширява анализа до четири модела и прагове до 0,95 в областта земеделие, като показва зависима от модела чувствителност според CPS, T-CPS и Balance Score, подкрепена със сдвоено статистическо тестване спрямо базовата конфигурация. Фаза IV повтаря този анализ в областта на биоразнообразието при стандартизирана хардуерна среда и показва по-ниски най-добри прагове, по-голяма големина на подобренията и пълно съвпадение между CPS, T-CPS и Balance Score. В обобщение, резултатите показват, че ефектите от прага зависят както от модела, така и от тематичната област, и че праговете с най-добри резултати изискват настройване спрямо конкретния корпус. Във всички фази компонентите на Obj.1 - процедура за оценка, инфраструктура и експериментален дизайн - са прилагани последователно, с което е изпълнена Obj.4 и е осигурена емпирична подкрепа за D1, D2 и D3.

ПЕТА ГЛАВА: ДИСКУСИЯ И НАСОКИ ЗА БЪДЕЩА РАБОТА

Дисертационният труд се ръководи от една изследователска цел и четири задачи. Целта е да се разработи рамка за оценка на RAG, която подпомага вземането на решения за конфигуриране на извличането в RAG системи с LLM с отворен код, с особен акцент върху настройването на прага на сходство. Obj. 1 дефинира и реализира основните компоненти на рамката: a) процедура за оценка, отчитаща прага на сходство, b) платформата PaSSER като инфраструктура за възпроизводимост и c) контролиран експериментален дизайн. Obj. 2 и 3 определят критериите за подбор на модели, както и метриките и процедурите за тяхното изчисляване. Obj. 4 прилага тези основи чрез контролирано тестване и анализ при различни модели и тематични области. Първа глава очертава изследователските дефицити. Втора и Трета глава представят платформата, обосновават избора на модели, метриките и процедурите за изчисляване. Четвърта глава представя експерименталните резултати. По-нататък са обобщени основните изводи: Секция 5.1 разглежда изследователските въпроси, Секция 5.2 представя научно-приложните приноси, Секция 5.3 обсъжда ограниченията, а Секция 5.4 очертава бъдещи насоки за изследване.

5.1 Отговори на изследователските въпроси

Секции 5.1.1–5.1.3 обобщават експерименталните доказателства от Фази II–IV по трите поставени въпроса.

5.1.1 (RQ1): Води ли промяната на прага на сходство до измерими промени в качеството на генерираното съдържание?

Конфигурацията на прага на сходство води до статистически значими ефекти върху качеството на генериране. В корпусите за земеделие и биоразнообразие варирането на прага в диапазона 0,50–0,95 дава CPS подобрения до +4,58% (Фаза III, земеделие; Таблица 4.7) и +13,32% (Фаза IV, биоразнообразие; Таблица 4.14) спрямо базовата линия.

Пилотните експерименти от Фаза II дават първи признак, че CPS се изменя според настройката на прага дори в сравнително тесен диапазон. При 101 двойки въпрос-отговор и три модела с 7B параметри е показано, че оптималните прагове се различават по модел: Mistral 7B и Llama 2 7B достигат най-добри резултати при 0,55, а Orca 2 7B - при 0,65. Тези резултати обосновават разширяването на диапазона на прага и на набора от модели в следващите фази.

Фаза III разширява анализа до четири модела, 369 въпроса и прагове в диапазона 0,50-0,95, като включва и статистическо тестване спрямо базовата линия. При тези условия Granite 3.2 8B и Mistral 7B v0.3 показват най-ясни признаци за систематична чувствителност към прага, със статистически значими положителни ефекти съответно при 7 от 10 и 6 от 10 прага. При DeepSeek R1 8B и Llama 3.1 8B значимите подобрения са по-малко, а ефектите от прага са по-ограничени или смесени.

Фаза IV прилага същата процедура за оценка към корпуса за биоразнообразие. Резултатите отново показват, че изборът на праг влияе върху качеството на генериране, но големината и посоката на тези ефекти се различават по модел. Mistral 7B v0.3 показва значимо подобрение при 9 от 10 прага, докато при Granite 3.2 8B широкият ефективен диапазон от Фаза III се свива до само 3 статистически значими прага във Фаза IV. DeepSeek R1 8B показва положителни ефекти при по-ниски прагове и влошаване при по-строги стойности, а при Llama 3.1 8B преобладават отрицателните ефекти от прага. Във Фаза IV колебанията в качеството на изхода нарастват при всички модели, което се отразява в по-високи стойности на коефициента на вариация.

Като цяло, тези резултати дават положителен отговор на RQ1: конфигурацията на прага на сходство води до измерими и в много случаи статистически значими ефекти върху качеството на генериране. Подобрения до +4,58% в областта земеделие и до +13,32% в областта биоразнообразие са постигнати само чрез калибриране на прага, без повторно обучение на модела или архитектурни промени. Оптималният праг обаче не е постоянен при различните модели и тематични области, което показва, че изборът му трябва да се установява експериментално за конкретната конфигурация, а не да се приема предварително.

5.1.2 (RQ2): Различава ли се ефектът от прага на сходство при различните езикови модели?

Наблюдавана е съществена модело-зависима вариация в чувствителността към прага. Моделите се различават по оптимални стойности на прага, широчина на ефективния диапазон, степен на CPS подобрение и устойчивост на резултатите между конфигурациите.

Mistral 7B v0.3 показва най-широка и най-устойчива чувствителност към прага. Във Фаза III (земеделие) значими подобрения са наблюдавани при 6/10 прага в обхвата 0,55–0,95. Във Фаза IV (биоразнообразие) чувствителността нараства до 9/10 значими прага, с максимално подобрение +13,32% CPS при праг 0,80. Сред четирите модела, оценявани във Фази III–IV, Mistral 7B v0.3 е единственият, запазващ широка положителна чувствителност и в двата домейна.

Granite 3.2 8B показва плато-поведение във Фаза III (7/10 значими прага в обхвата 0,65–0,95), но се свива до 3/10 значими прага във Фаза IV. Това свиване е най-голямата междудомейнна промяна в широчината на чувствителността сред оценяваните модели, макар усложняващите фактори на Фаза IV (Секция 5.3.2) да затрудняват тълкуването.

Llama 3.1 8B показва ограничена положителна чувствителност във Фаза III и предимно отрицателни ефекти във Фаза IV (6/10 значими прага: 5 отрицателни, 1 положителен). Тази инверсия предполага, че калибрирането на прага за Llama 3.1 8B е особено чувствително към характеристиките на корпуса или експерименталните условия.

DeepSeek R1 8B показва ограничена чувствителност във Фаза III (2/10 значими прага), но по-силни ефекти във Фаза IV с двуполюсно поведение: значими подобрения при ниски прагове (0,55–0,60)

и значимо влошаване при високи (0,90–0,95). Ниските стойности на CV (0,129–0,166) допринасят за най-високия Balance Score сред всички комбинации модел-праг във Фаза IV (0,678 при праг 0,55; Таблица 4.21), което показва, че при прилагане на оценяване, претеглено за стабилност, DeepSeek R1 8B може да бъде конкурентен въпреки ограниченото средно CPS подобрение.

Тези резултати дават отговор на RQ2: чувствителността към прага се различава съществено между моделите по степен и характер. Чувствителността варира от широка и междудомейнно стабилна (Mistral 7B v0.3), до тясна и домейн-чувствителна (Granite 3.2 8B), инвертирана между домейните (Llama 3.1 8B) и двуполюсна с висока устойчивост на резултатите (DeepSeek R1 8B). Тези разлики показват, че изборът на праг не може да бъде обобщен между моделните архитектури; конфигурации, ефективни за един модел, могат да влошат производителността на друг.

5.1.3 (RQ3): Валидни ли са сходни диапазони на прага при различен домейн?

Чувствителността към прага е сравнена между корпус за земеделие (Фаза III) и корпус за биоразнообразие (Фаза IV). Наблюдавани са систематични разлики по множество показатели, макар тълкуването да е усложнено от едновременните промени в процедурата за генериране на въпроси, хардуерната конфигурация и характеристиките на корпуса (Секция 5.3.2).

Оптималните прагове се изместват систематично надолу във Фаза IV. И четирите модела постигат максимален CPS при по-ниски прагове в домейна биоразнообразие, отколкото в земеделието, с изместване от −0,05 (Llama 3.1 8B) до −0,35 (DeepSeek R1 8B; Таблица 4.20). Степента на подобрение се различава значително: Mistral 7B v0.3 (+13,32% срещу +4,58%), Granite 3.2 8B (+6,95% срещу +1,25%) и DeepSeek R1 8B (+8,45% срещу +1,01%) постигат по-големи CPS подобрения във Фаза IV, докато Llama 3.1 8B показва сходна степен при предимно отрицателни ефекти. Класирането по CPS подобрение се променя: DeepSeek R1 8B се изкачва от четвърто на второ място, а Llama 3.1 8B пада от второ на четвърто. Корелационните структури на ниво метрики също се изместват между фазите (Фигури 4.3 и 4.7), което указва чувствителност на взаимовръзките между метриките към корпуса.

5.2 Научно-приложни приноси

Три научно-приложни приноса, организирани като компоненти от интегрираната рамка за оценка, адресират дефицитите, идентифицирани във увода: процедура за оценка с отчитане на прага и композитни показатели (C1, компонент Процес на оценяване, адресиращ D1), инфраструктура за възпроизводимост с блокчейн-базиран запис (C2, компонент Инфраструктура, адресиращ D2) и сравнителни емпирични насоки за избор на модел и праг (C3, компонент Емпирични резултати, адресиращ D3). Секции 5.2.1–5.2.3 характеризират всеки принос.

5.2.1 Процедура за оценка

Секция 1.4 идентифицира, че съществуващите рамки за оценка, като RAGAS, RGB, TREC RAG Track и TruLens, обикновено оценяват RAG системите при фиксирани настройки на извличането и не разглеждат прага на сходство като самостоятелна експериментална променлива. В много реализации изборът на праг остава конфигурационно решение, често евристично и прилагано без систематична обосновка. Съществуващите подходи обикновено поставят акцент върху средното представяне и рядко оценяват как качеството на изхода се променя между отделните заявки.

За преодоляване на този дефицит в настоящото изследване е разработена процедура за оценка, отчитаща прага на сходство, при която той се разглежда като независима променлива, а поведението на системата се оценява в диапазон на селективност на извличането, вместо при една фиксирана настройка. Тази процедура е реализирана чрез платформата PaSSER и е приложена във Фази II-IV. Фаза II въвежда пилотно изследване на прага в диапазона 0,50-0,80. Фаза III разширява диапазона до 0,95 и включва статистическо тестване в областта земеделие. Фаза IV прилага същата процедура към областта на биоразнообразието, за да се изследват разликите между тематичните области.

Резултатите показват, че ефектите от прага са съществени и не могат да бъдат сведени до една фиксирана стойност по подразбиране. Във Фаза III праговете с най-добри резултати по CPS попадат в диапазона 0,90-0,95, докато критериите, отчитащи стабилността, разширяват ефективния диапазон до 0,65-0,95. Във Фаза IV праговете с най-добри резултати се изместват надолу до 0,55-0,85. Разликите в представянето между отделните прагове достигат до +4,58% подобрение по CPS в земеделието и +13,32% в биоразнообразието, което показва, че оценяването при една-единствена стойност на прага може да не представя реалното поведение на системата в целия диапазон на извличане.

Този принос включва и три формулировки за композитно оценяване. CPS обединява

разнородни метрики чрез претеглено сумиране след нормализация, като използва панел от 9 метрики, разпределени между лексикално припокриване (30%), семантично сходство (25%), гладкост и точност (25%) и езиково моделиране (20%). T-CPS разширява CPS чрез включване на коефициента на вариация (CV), като по този начин отчита стабилността на изхода между отделните заявки. Balance Score допълнително оценява връзката между подобрението и колебанията.

Във Фази III и IV T-CPS и CPS показват много висока корелация ($\rho > 0,99$), което означава, че в изследваните конфигурации средното представяне остава водещият сигнал за подреждане. Това не прави T-CPS излишен. Напротив, показва, че оценката, отчитаща стабилността, обикновено съвпада със средното представяне, като същевременно позволява да се открояят случаи, при които предпочитаният праг се променя, след като се отчетат колебанията. Най-ясният пример е DeepSeek R1 8B във Фаза III, при който CPS достига максимум при 0,90, а T-CPS предпочита 0,65, тъй като по-ниските колебания изместват оптимума, отчитащ стабилността.

Този принос адресира D1, като заменя оценяването при фиксиран праг с процедура, отчитаща прага на сходство и обхващаща както равнището на представяне, така и неговата стабилност. Представените резултати имат изследователски характер, а потвърдителни изследвания с по-строг статистически контрол биха засилили тяхното обобщение.

5.2.2 Инфраструктура за възпроизводимост

Вторият принос е платформата PaSSER, разработена с цел да отговори на проблемите с възпроизводимостта при оценяването на RAG системи. PaSSER предоставя среда с достъп през браузър за контролирано експериментиране, като обединява конфигурирането на извличането, изпълнението на модела, изчисляването на метриките и експортирането на резултатите в единен работен процес.

Ключов елемент на тази инфраструктура е проследяването на произхода чрез блокчейн. За всеки експериментален цикъл платформата записва върху блокчейна Antelope метриките за оценка, данните за времето на изпълнение и идентификаторите на изпълнението, а бекендът съхранява свързания конфигурационен контекст, включително идентификатора на модела, параметрите за извличане, настройките за декодиране и идентификаторите на използваните набори от данни. Тази архитектура осигурява неизменяеми записи, проверими времеви маркери и възможност за последващо извличане на метаданните за експериментите с цел независима проверка.

Записът в блокчейн не гарантира точно възпроизвеждане на цялата изчислителна среда. Той обаче подобрява възпроизводимостта, като съхранява експерименталните условия и получените резултати в проверима форма. По този начин приносът адресира D2 не чрез претенция за пълна възпроизводимост, която би била необоснована, а чрез повишаване на прозрачността, проследимостта и възможността за последваща верификация на експерименталните изпълнения.

Материалите, свързани с дисертационния труд, са публично достъпни на две места в GitHub. Архивът на дисертацията, организиран по фази и включващ експерименталните резултати и съпътстващите изследователски материали, е достъпен на https://github.com/M33rschaum/passer-thesis-archive. Актуалните реализации на PaSSER, включително оригиналната платформа PaSSER и свързаните с нея хранилища, като maPaSSER и PaSSER-SR, са достъпни чрез GitHub на платформата SCPDx на адрес https://github.com/scpdxtest.

5.2.3 Практически насоки за избор на модел при внедряване на RAG

Третият принос се състои в извеждането на сравнителни емпирични насоки за подбор и конфигуриране на LLM с отворен код при RAG условия. Този принос адресира липсата на систематични данни за вариантите за внедряване с отворен код, особено когато изборът на модел и конфигурацията на извличането трябва да се правят при ограничения, свързани със сигурността, разходите или инфраструктурата.

Въз основа на критериите за подбор на модели и процедурите за изчисляване на метриките, дефинирани в Трета глава, както и на контролираните експерименти, представени в Четвърта глава, в дисертационния труд са проведени над 38 000 оценки в рамките на четири фази, две тематични области и няколко хардуерни конфигурации. Тези резултати правят чувствителността към прага видима и съпоставима при вариантите за внедряване в диапазона 7-8 милиарда параметри.

Четири извода са особено важни от практическа гледна точка. Първо, моделите се различават съществено по профила си на реакция към прага: някои показват широк и устойчив положителен ефект, докато при други той е тесен, нестабилен или двуполюсен. Второ, ефективните диапазони на прага,

установени във Фази III и IV, дават отправни точки за калибриране, вместо да се разчита на общи стойности по подразбиране. Трето, промените между тематичните области показват, че настройките на прага трябва да се преразглеждат при промяна на корпуса. Четвърто, Balance Score дава допълнителна основа за вземане на решения при внедряване, когато наред със средното представяне има значение и последователността на резултатите.

Наблюдаваните ефекти на прага са особено важни по отношение на две проблемни точки при извличането, описани в предходни изследвания [47]: изключване на съответстващи източници при прекалено строг филтър и допускане на слабо свързан материал при прекалено либерален филтър. Настоящите резултати показват, че чувствителността на моделите към тези проблеми се различава между отделните системи, макар че причинните механизми не са изолирани в рамките на този дисертационен труд и биха изисквали самостоятелни аблационни изследвания.

Този принос адресира D3, като предоставя емпирични насоки за калибриране на прага и сравнение между модели при RAG сценарии с отворен код, отчитайки както средното качество, така и стабилността при промяна на корпуса.

5.3 Ограничения

Няколко фактора ограничават обобщаването на резултатите извън оценяваните корпуси, модели и конфигурации.

5.3.1 Ограничение на обхвата

Оценени са два домейна: земеделие (Фаза III) и биоразнообразие (Фаза IV). И двата съдържат техническо и нормативно съдържание с формален език и добре дефинирана терминология. Дали подобни зависимости на чувствителността към прага се проявяват при други типове домейни не е проверено. Групата модели е ограничена до модели с отворен код в диапазона 7–8B параметри, за да се запази възможността за ползване на хардуер от среден клас. Чувствителността към прага може да се различава извън този диапазон.

5.3.2 Ограничения на експерименталния дизайн

Експериментите са проведени на разнороден хардуер поради практически ограничения. Фаза III използва M1, M2 и CPU-only среди с контекстни буфери (2 048–10 000 токена), докато Фаза IV стандартизира до M1 хардуер с фиксиран контекстен буфер от 16 000 токена, елиминирайки хардуерната/контекстна вариация, но въвеждайки едновременни промени в домейна и метода за генериране на въпроси.

Векторните хранилища са създадени с вграждания от Mistral 7B чрез Ollama, с фиксирано фрагментиране (1 024 знака, припокриване 50) във всички фази. Използването на един модел за векторизация осигурява последователни условия за извличане, но може да повлияе на резултатите поради съвместимостта между векторния и генеративния компонент. Тъй като Mistral 7B служи едновременно като модел за вграждания и като един от оценяваните модели за генериране, съществува потенциален усложняващ фактор; изолирането на този ефект би изисквало оценяване с независими модели за вграждания (Секция 5.4.3). Въпросите за земеделие са генерирани с Mistral 7B (Фази I–III), а за Фаза IV е използван Claude Opus, което осигурява частична кръстосана проверка, но не изолира ефектите на вгражданията.

Сравнението между домейнитевъв Фази III и IV се различава в няколко отношения освен съдържанието на домейна (модел за генериране на въпроси, хардуерна стандартизация, характеристики на корпуса). Тези едновременни промени не позволяват приписване на наблюдаваните изменения на прага конкретно на свойствата на домейна; данните следва да се тълкуват като предварителни доказателства, предполагащи, че може да е необходимо домейн-специфично калибриране.

5.3.3 Ограничения при измерванията и анализа

Метриките за perplexity (Laplace и Lidstone) са изчислени чрез NLTK n-грамни езикови модели, а не чрез вероятности на ниво токен от трансформер (Секция 3.4.3), и функционират като модел-независими индикатори. Наборът B-RT използва [CLS]-базирани проекции и не е валидиран спрямо човешки преценки (Секция 3.4.5). Фази III и IV докладват p-стойности без корекция, за да запазят статистическата коректност (Секция 3.5.3). Всяка фаза включва 40 сравнения спрямо базовата линия; при $\alpha = 0{,}05$ приблизително 2 значими резултата на фаза биха могли да се появят случайно при нулева хипотеза. Читателите следва да тълкуват цялостната картина на резултатите, а не да разчитат на

отделни p-стойности изолирано. Резултатите са с проучвателен характер и мотивират по-нататъшно изследване, а не окончателни предписания за прагове.

5.3.4 Причинно-следствена интерпретация

Експерименталният дизайн установява емпирични зависимости между конфигурацията на прага на сходство и качеството на генериране, включително модел-зависими оптимуми (0,55–0,95), домейн-зависими изместванения между земеделие и биоразнообразие и баланс стабилност-производителност, оценен чрез Balance Score. Причинните механизми, стоящи зад тези модели, обаче не са експериментално изолирани. Множество фактори варират между условията — характеристики на корпуса, формулиране на въпроси, хардуерна конфигурация и архитектура на модела — което не позволява приписване на наблюдаваните ефекти на единичен механизъм. Потвърждаването на причинно-следствени обяснения би изисквало контролирани аблационни изследвания, при които се променя един фактор, а останалите се поддържат постоянни.

5.4 Насоки за бъдеща работа

Бъдещата работа е организирана в три направления: разширяване на обхвата, процедурни разширения и валидиране/проверка.

5.4.1 Разширяване на обхвата

Разширяване на анализа на чувствителността към прага към допълнителни домейни извън земеделието и биоразнообразието и към модели извън диапазона 7–8B параметри. Контролирано вариране на домейна, при което се променя само корпусът, би изолирало домейн ефекти от усложняващите фактори, документирани в Секция 5.3.2. Покритието на модели може да бъде разширено към по-малки и по-големи модели, дообучени варианти и, където е възможно, затворени модели при съпоставими условия.

5.4.2 Процедурни разширения

Изследване на адаптивен подбор на прага, при който селективността на извличането варира според характеристиките на въпроса. Разширяване на PaSSER за записване на информация за произхода на ниво извличане (извлечени пасажи и стойности на сходство) и интегриране на IPFS за адресирано по съдържание съхранение на корпуси и набори от данни. Характеризиране на баланса ефективност-качество чрез измерване на латентност, използване на паметта и производителност при различни прагове.

5.4.3 Валидиране и проверка

Провеждане на валидиращи изследвания с участието на хора, за да се оцени дали праговете, подбрани чрез CPS/T-CPS, съответстват на подобрения, възприемани от хора, и за валидиране на B-RT спрямо човешки преценки. Провеждане на контролирани аблационни изследвания за изолиране на ефектите на вграждания, фрагментирането, генерирането на въпроси, хардуера и самостоятелните промени в домейна. Проучване на алтернативни подходи за измерване, включително perplexity, базирана на трансформер, когато са достъпни logits, и прилагане на корекции за множествени сравнения (напр. Bonferroni или Benjamini-Hochberg FDR) в потвърдителни изследвания.

5.5 Пета глава: Обобщение

Експерименталните резултати са обобщени по трите поставени въпроса. Три научно-приложни приноса са характеризирани и съпоставени с дефицитите, идентифицирани в Първа глава. Документирани са ограниченията на обхвата, експерименталния дизайн, измерванията и анализа, както и границите при тълкуване. Бъдещата работа е организирана в разширяване на обхвата, процедурни разширения и усилия за валидиране и проверка, адресиращи документираните ограничения.

ЗАКЛЮЧЕНИЕ - РЕЗЮМЕ НА ПОЛУЧЕНИТЕ РЕЗУЛТАТИ

Разработена е Рамка за оценка на генериране, подпомогнато чрез външно извличане (RAG), която обединява три компонента: (1) компонент "Процедура за оценка", които отчита прага на сходство, (2) инфраструктурен компонент с инструменти за повторяеми експерименти и данни за проследимост, и (3) компонент "Емпирични данни" за сравнение между модели и домейни. Рамката преодолява три дефицита в практиката: липса на систематично оценяване при вариране на прага, недостатъчни средства за повторяемост и проследимост на експериментите и липса на практически насоки за избор на модел и праг при решения с отворен код и локална инфраструктура.

Проектирана и реализирана е платформата PaSSER като уеб-приложение с отворен код, достъпно през браузър. Поддържано е тестване на параметри за извличане, многометрично оценяване с композитни показатели и запис на данни за проследимост в блокчейн чрез регистъра Antelope.

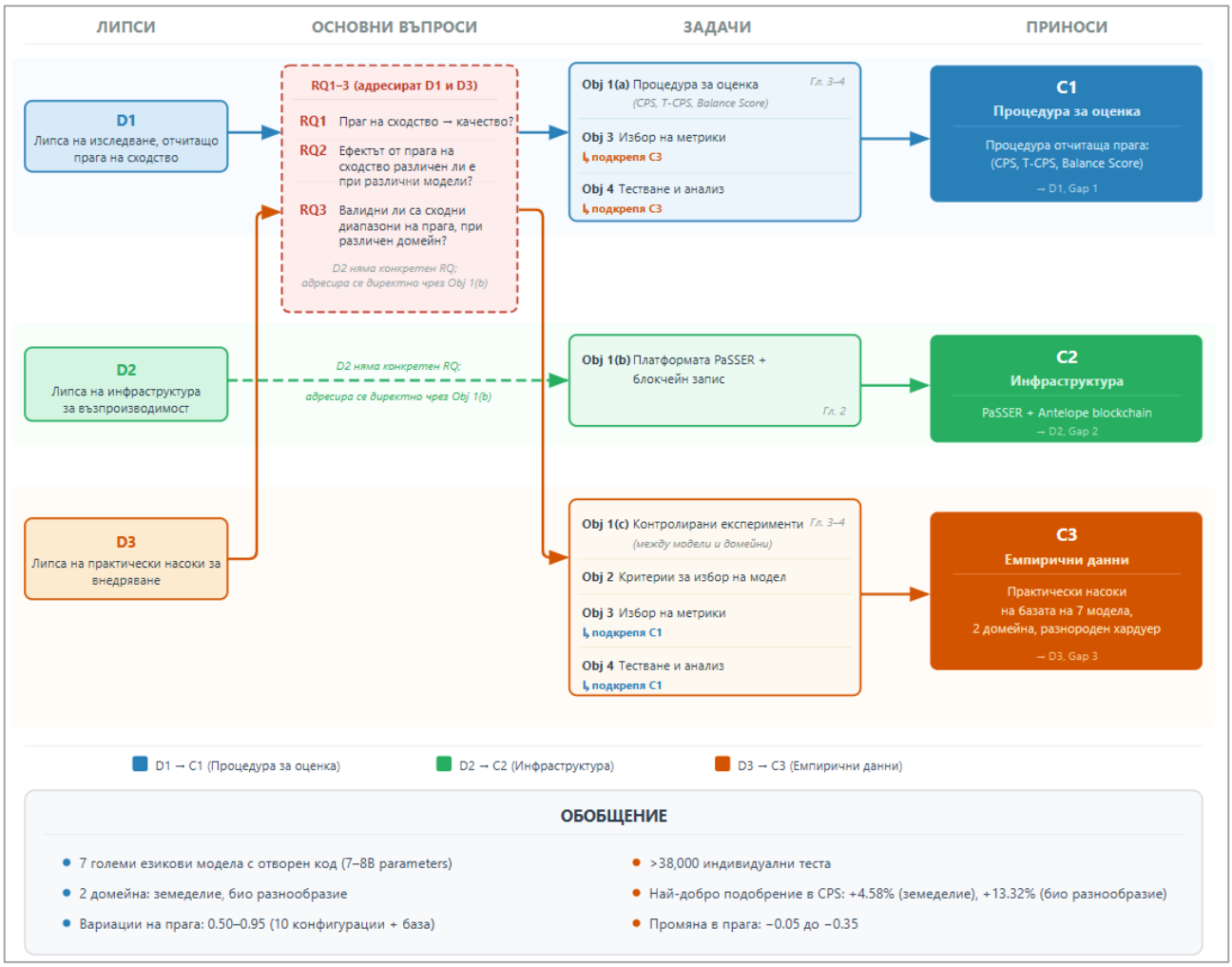
Разработени са три инструмента за композитно оценяване. Композитният показател (CPS) дава единна база за сравнение между конфигурации на прага. Праговият композитен показател (T-CPS) добавя оценяване на последователността на изхода чрез структура награда-наказание, базирана на коефициент на вариация (CV), използван тук като мярка за нестабилност. Balance Score измерва компромиса между качество и стабилност.

Оценени са седем езикови модела с отворен код в диапазона 7-8 милиарда параметъра в четири експериментални фази, два домейна (земеделие и биоразнообразие) и над 38 000 индивидуални оценки. Фаза I потвърждава работата на платформата "от край до край" при фиксирано top-k извличане. Фаза II въвежда оценяване, което отчита прага, и показва, че чувствителността към прага е модел-зависима. Фаза III разширява анализа до четири по-нови модела и диапазон 0,50-0,95 със статистическа проверка. В домейна земеделие са отчетени CPS подобрения до 4,58% и два типа поведение: широки области с подобрение и тесни диапазони, в които прагът работи добре. Фаза IV повтаря оценяването в домейна биоразнообразие, където ефектите от прага са значително по-големи и пиковите CPS подобрения достигат 13,32%. При преминаване от земеделие към биоразнообразие най-добрите прагови конфигурации се изместват надолу за всичките четири модела, от -0,05 (Llama 3.1 8B) до -0,35 (DeepSeek R1 8B). Нестабилността на изхода нараства с 67-105% (CV) спрямо земеделието, което потвърждава, че чувствителността към прага зависи от домейна и не е само свойство на модела. Сравнени са CPS и T-CPS и е показано, че оценяване по средна производителност и оценяване, което отчита последователността, може да доведе до различни препоръки за праг. При два от четирите модела прагът с най-добър резултат по CPS не съвпада с този по T-CPS.

Покрити са и четирите задачи на изследването. Obj.1 е реализирана чрез трите компонента на рамката: инфраструктурният компонент с данни за проследимост в блокчейн (b, Втора глава), компонентът за оценяване с отчитане на прага чрез композитни показатели (a, Трета и Четвърта глава) и контролираният експериментален дизайн, който дава сравними резултати между модели и домейни (c, Четвърта глава). Obj.2 е изпълнена чрез критерии за избор на модели според практическа приложимост, лицензиране и изчислителни изисквания (Трета глава). Obj.3 е изпълнена чрез дефиниране и последователно прилагане на процедури за изчисление на метрики в пет направления на оценяване, с обобщаване в трите композитни инструмента (Трета глава). Obj.4 е постигната чрез контролирани експерименти в четири фази, два домейна и над 38 000 оценки при систематично вариране на прага (Четвърта глава)).

От емпиричните резултати следват три практически извода. Първо, настройването на прага дава измерими и статистически значими подобрения, но най-добрите стойности зависят едновременно от модела и от домейна, тоест една настройка не работи еднакво добре навсякъде. Второ, оценяване, което отчита последователността, е за предпочитане пред оценяване само по средни стойности, защото ограничава избора на конфигурации с високи резултати, но висока нестабилност. Трето, оценяването при вариране на прага трябва да се повтаря при смяна на домейна, тъй като най-добрите прагове се изместват надолу за всичките четири тествани модела при преминаване от земеделие към биоразнообразие.

Получени са три научно-приложни приноса, които заедно оформят Рамката за оценка. Проследимостта от дефицитите, през основните въпроси и задачите до приносите е показана на Фигура С.1.



Фигура С.1..Карта на проследимостта

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

[15] I. Radeva, I. Popchev, and M. Dimitrova, „Similarity thresholds in retrieval-augmented generation,“ in 2024 IEEE 12th Int. Conf. on Intelligent Systems (IS), Aug. 2024, pp. 1–7, doi: 10.1109/IS61756.2024.10705214. Тази работа подкрепя формулировката на CPS и анализа на чувствителността към прагове, представен в Глава 4 (Фаза II).

[16] M. Dimitrova, I. Popchev, and I. Radeva, „PaSSER: A platform for evaluating LLMs in RAG,“ in 2025 IEEE BdKCSE, 2025, p. 7, doi: 10.1109/BdKCSE67969.2025.11300500. Тази работа описва архитектурата и функционалностите на платформата PaSSER, разработени в Глава 2.

[17] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, „Web application for retrieval-augmented generation: Implementation and testing,“ Electronics, vol. 13, no. 7, p. 1361, Apr. 2024, doi: 10.3390/electronics13071361. Тази работа представя платформата PaSSER и метриките, обсъдени в Глави 2 и 3.

[18] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, „Multi-agent coordination strategies vs. retrieval-augmented generation in LLMs: A comparative evaluation,“ Electronics, vol. 14, no. 24, p. 4883, Jan. 2025, doi: 10.3390/electronics14244883. Тази работа документира T-CPS и Balance Score, докладвани в Глава 4 (Фаза IV).

[20] M. Dimitrova, „Retrieval-augmented generation (RAG): Advances and challenges,“ Problems of Engineering Cybernetics and Robotics (PECR), vol. 83, Jul. 2025, doi: 10.7546/PECR.83.25.03. Тази работа предоставя литературния обзор на RAG и анализа на рамки, формиращи основата на Глава 1.

Публикации [17] и [18] са индексирани в JCR-IF (Web of Science) и SJR (Scopus). Конферентни доклади [15] и [16] са индексирани в IEEE Xplore Digital Library. Статия [20] е публикувана от Издателство на Българската академия на науките „Проф. Марин Дринов“.

ЗАБЕЛЯЗАНИ ЦИТИРАНИЯ

[1] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, „Web Application for Retrieval-Augmented Generation: Implementation and Testing,“ Electronics, vol. 13, no. 7, 2024, doi:10.3390/electronics13071361.

Цитирана от:

[1.1] H. Andersson, "Retrieval-augmented generation with Azure OpenAI," M.S. thesis, Malardalen Univ., 2024.

[1.2] S. D’Urso, B. Martini, and F. Sciarrone, "A Novel LLM Architecture for Intelligent System Configuration," in Proc. Int. Conf. Information Visualisation (IV), Coimbra, Portugal, 2024, pp. 326-331, doi:10.1109/IV64223.2024.00063.

[1.3] D. Firdaus, I. Sumardi, and Y. Kulsum, "Integrating Retrieval-Augmented Generation With Large Language Model Mistral 7B for Indonesian Medical Herb," JISKA, vol. 9, no. 3, pp. 230-243, 2024, doi:10.14421/jiska.2024.9.3.230-243.

[1.4] H. Zhang, Z. Li, F. Liu, Y. He, Z. Cao, and Y. Zheng, "Design and Implementation of LangChain-based Chatbot," in Proc. Int. Seminar on AI, Computer Technology and Control Engineering (ACTCE), Wuhan, China, 2024, pp. 226-229, doi:10.1109/ACTCE65085.2024.00053.

[1.5] J. G. Ongris, E. Tjitrahardja, F. Darari, and F. J. Ekaputra, "Towards an Open NLI LLM-based System for KGs: A Case Study of Wikidata," in Proc. 7th Int. Seminar on Research of IT and Intelligent Systems (ISRITI), 2024, pp. 44-49, doi:10.1109/ISRITI64779.2024.10963661.

[1.6] C. K. Kitengera and M. K. Kasambya, "Développement d’une plateforme web d’évaluation des enseignements...," Revue Internationale Multidisciplinaire Etincelle, vol. 25, no. 2, pp. 1-22, 2024, doi:10.61532/rime252117.

[1.7] B. Lu, "Evaluating LLMs on large contexts: a RAG approach on text comprehension," Master’s thesis,

Univ. de Liège, 2024.
[1.8] M. M. Li, I. Nikishina, O. Sevgili, and M. Semman, "Wiping out the limitations of Large Language Models: A Taxonomy for Retrieval Augmented Generation," arXiv:2408.02854, 2024, doi:10.48550/arXiv.2408.02854.
[1.9] M. Olaosegba, "Next-Gen AI Optimization Tools for AWS Cloud Cost Control," IJFMR, 2024.
[1.10] P. Phukon, Y. Lokhar, and P. P. Ray, "Localized Open-Source LLM Aware RAG of Legal Documents...," in Proc. Int. BIT Conf. (BITCON), 2024, pp. 1-6, doi:10.1109/BITCON63716.2024.10985396.
[1.11] S. Rani, S. G. Deepika, D. Devdharshini, and H. Ravindran, "Augmenting Code Sequencing with RAG...," in Proc. SSITCON, 2024, pp. 1-7, doi:10.1109/SSITCON62437.2024.10796587.
[1.12] S. Dudhmande et al., "Textual Compression Using Lamini-LM," IRJAEM, vol. 2, no. 5, pp. 1536-1540, 2024, doi:10.47392/IRJAEM.2024.0208.
[1.13] S. Bouzid and L. Piron, "Leveraging Generative AI in Short Document Indexing," Electronics, vol. 13, no. 17, 2024, doi:10.3390/electronics13173563.
[1.14] K. Traykov, "A Framework for Security Testing of Large Language Models," in Proc. 12th IEEE Int. Conf. on Intelligent Systems (IS), Varna, Bulgaria, 2024, pp. 1-7, doi:10.1109/IS61756.2024.10705238.
[1.15] L. Werkman, "Assessing the potential of leveraging LLaMA-2...," thesis, Luleå Univ. of Technology, 2024.
[1.16] W. Wilmi and N. Roslund, "Implementering av RAG för automatiserad analys av hållbarhetsrapportering...," thesis, KTH Royal Inst. of Technology, 2024.
[1.17] Y. Xu et al., "Development of an Enterprise Knowledge Base System Based on Elasticsearch," in Proc. ISPCEM, 2024, pp. 186-190, doi:10.1109/ISPCEM64498.2024.00039.
[1.18] Y. Song, Enhancing Classroom Dialogue Productiveness: Exploring the Potential of Artificial Intelligence. London, U.K.: Routledge, 2024, doi:10.4324/9781003543039.
[1.19] Z. Zhong et al., "Mix-of-Granularity: Optimize the Chunking Granularity for RAG," arXiv:2406.00456, 2024.
[1.20] J. O. Agada et al., "A Systematic Review of Key RAG Systems...," arXiv:2507.18910, 2025, doi:10.48550/arXiv.2507.18910.
[1.21] A. Guyyala et al., "RAG-based AI Agents for Multilingual Help Desks...," Int. J. Computer Applications, vol. 187, no. 56, pp. 15-28, 2025, doi:10.5120/ijca2025925964.
[1.22] O. Barcelos et al., "Technological Convergence Identification Model (TCIM)...," Revista E-TECH, vol. 18, no. 1, 2025, doi:10.18624/e-tech.v18i1.1444.
[1.23] C. Yu et al., "Safety Devolution in AI Agents," 2025, doi:10.48550/arXiv.2505.14215.
[1.24] D. Costa et al., "Mycroft: Retrieval Augmented Generation for SDK Documentation," in Proc. NATL, 2025, doi:10.5121/csit.2025.152211.
[1.25] R. Dayarathne et al., "Comparing the Performance of LLMs in RAG-Based QA...," in AI in Education Technologies, LNDECT, vol. 228. Singapore: Springer, 2025, doi:10.1007/978-981-97-9255-9_26.
[1.26] E. H. Omoush et al., "Advancing Arabic Medical QA Systems with RAG...," in Proc. ICTCS, 2025, pp. 511-516, doi:10.1109/ICTCS65341.2025.10989446.
[1.27] Y. Fan et al., "Research on the Online Update Method for RAG Model...," in Proc. NNICE, 2025, pp. 1740-1744, doi:10.1109/NNICE64954.2025.11063821.
[1.28] F. Shen et al., "Development of a Convenient Accounting System Based on SpringBoot+Vue," in Proc. CITSC, 2025, pp. 167-171, doi:10.1109/CITSC64390.2025.00038.
[1.29] F. Ehrlich-Sommer et al., "ForestGPT and Beyond...," Electronics, vol. 14, no. 18, p. 3583, 2025, doi:10.3390/electronics14183583.
[1.30] H. Mahfoud et al., "AI Chatbots for Healthcare Maintenance...," TQM J., 2025, doi:10.1108/TQM-10-2024-0394.
[1.31] G. Iieva and G. A. Tsihrintzis, "Editorial Note to Special Issue...," Electronics, vol. 14, no. 10, p. 1925, 2025, doi:10.3390/electronics14101925.

[1.32] L. A. Sanjani et al., "Performance Analysis of LLM Models with RAG and Fine-Tuning T5...," in Proc. ICoCSETI, 2025, pp. 152-157, doi:10.1109/ICoCSETI63724.2025.11018908.
[1.33] B. T. Mahardika and A. M. Hasan, "Application of GPT in Chatbots...," Eduvest, vol. 5, no. 6, pp. 6235-6247, 2025, doi:10.59188/eduvest.v5i6.51321.
[1.34] N. A. Akbar et al., "Novel Approach for Leveraging Agent-Based Experts...," in AIxIA 2024, LNCS, vol. 15450. Cham, Switzerland: Springer, 2025, doi:10.1007/978-3-031-80607-0_2.
[1.35] B. M. Praneeth et al., "Optimization of Customer Feedback Summarization...," IEEE Access, vol. 13, pp. 124319-124332, 2025, doi:10.1109/ACCESS.2025.3588337.
[1.36] P. Pany, "Reasoning Engine with Pre-Trained LLMs: An Operation GPT," IJRASET, vol. 13, no. 4, pp. 2452-2463, 2025, doi:10.22214/ijraset.2025.68761.
[1.37] S. K. Mahjour and S. S. Mahjour, "Intelligent Reservoir Decision Support...," 2025, doi:10.48550/arXiv.2509.11376.
[1.38] S. Chen et al., "Customized large-scale model for human-AI collaborative operation...," Appl. Energy, vol. 393, pp. 126-169, 2025, doi:10.1016/j.apenergy.2025.126169.
[1.39] T. Jung and I. Joe, "An Intelligent Docent System with a Small Language Model (sLLM) Based on RAG," Appl. Sci., vol. 15, no. 17, p. 9398, 2025, doi:10.3390/app15179398.
[1.40] C.-N. Tirpescu and E. Velescu, "Enhancing Veterinary Education...," Procedia Comput. Sci., vol. 270, pp. 3828-3837, 2025, doi:10.1016/j.procs.2025.09.508.
[1.41] W. Ke et al., "Large Language Models in Document Intelligence: A Comprehensive Survey...," ACM Trans. Inf. Syst., vol. 44, no. 1, 2025, doi:10.1145/3768156.
[1.42] A. J. Winata et al., "Utilizing Large Language Models for Developing Automatic Question Generation in Education," in Proc. ICADEIS, 2025, doi:10.1109/ICADEIS65852.2025.10933227.
[1.43] Y. Benitez-Morejon et al., "Question-Answering Systems for Tourism...," in MISNC 2025, CCIS, vol. 2729. Cham, Switzerland: Springer, doi:10.1007/978-3-032-09945-7_22.
[1.44] J. Qi, Mitigating Translation Hallucinations in Large Language Models: A Chain of Thought and RAG-Based Approach, Ph.D. research proposal, The Chinese Univ. of Hong Kong, 2024-2025.
[1.45] R. Kumar and Y. Qu, "Utilizing Large Language Model Enabled Agents to Streamline Business Decision Making," Eur. J. Electr. Eng. Comput. Sci., vol. 9, no. 5, pp. 14-21, Sep. 2025, doi:10.24018/ejece.2025.9.5.717.
[1.46] I. P. A. E. Pratama, I. M. O. Widyantara, Linawati, and N. Gunantara, "Bibliometric Analysis of AI-Based Prototype Proposal for User Security Awareness in Healthcare," JOIV, vol. 9, no. 3, pp. 982-994, May 2025, doi:10.62527/joiv.9.3.3319.
[1.47] S. Gandla, Automated Test Code Generation from Textual Descriptions Using Generative AI, Master's thesis, Blekinge Inst. of Technology, 2024.
[1.48] N. S. Patil, A. J. Koyande, A. V. Thakur, P. B. Kadam, and P. G. Moholkar, "RAG Chatbots: Implementing Large Language Models in Retrieval-Augmented Generations," in Smart Trends in Computing and Communications, LNNS, vol. 1363, pp. 401-410, 2025, doi:10.1007/978-981-96-2885-8_33.
[1.49] I. Balen, Sustav korisnicke podrske temeljen na bazi znanja..., Undergraduate thesis, Faculty of Electrical Engineering and Computing (FER), Univ. of Zagreb, Jun. 2024.
[1.50] Y. Jiao, S. Ouyang, M. Zhong, Y. Zhang, L. Ding, S. Zhou, and J. Han, "Retrieval and Structuring Augmented Generation with LLMs for Web Applications," in Companion Proc. ACM Web Conf. 2025 (WWW '25 Companion), pp. 25-28, May 2025, doi:10.1145/3701716.3715870.
[1.51] Z. Liu, Design and Implementation of an AI-based Agent to Inform Best Practices on Test Case Execution Routines, Master's thesis, Univ. of Zurich, Jun. 29, 2025, doi:10.5167/uzh-278942.

[2] I. Radeva, I. Popchev, and M. Dimitrova, „Similarity Thresholds in Retrieval-Augmented Generation," in Proc. 12th IEEE Int. Conf. on Intelligent Systems (IS'24), Varna, Bulgaria, Aug. 29-31, 2024, pp. 1-7,

doi:10.1109/IS61756.2024.10705214.

Цитирана от:

[2.1] D. Ayepah-Mensah et al., "A RAG-Assisted DRL Framework for Microservices Deployment in 6G Vehicular Networks," in Proc. WiMob 2025, Marrakesh, Morocco, 2025, pp. 1-6, doi:10.1109/WiMob66857.2025.11257559.

[2.2] Y. Bondalapati and H. N. BM, "Scalable RAG with Kubernetes for Enhanced Document Intelligence," in Proc. CICCS 2025, Bengaluru, India, 2025, pp. 1-6, doi:10.1109/CICCS66437.2025.11280266.

[2.3] A. Jadhav et al., "AI-Driven Diagnosis Predictive Chatbot for Healthcare," in Proc. WorldSUAS 2025, 2025, doi:10.1109/WorldSUAS66815.2025.11199219.

[2.4] J. Van Nooten et al., "One Size Does Not Fit All: Exploring Variable Thresholds for Distance-Based Multi-Label Text Classification," arXiv:2510.11160, 2025, doi:10.48550/arXiv.2510.11160.

[2.5] X. Sun, C. Liang, Q. Wang, et al., "Mesh RAG: Retrieval Augmentation for Autoregressive Mesh Generation," arXiv:2511.16807, 2025.

[2.6] K. Traykov and Y. Kolova, "Analysis of Methods for Evaluating Responses of LLMs in Retrieval-Augmented Generation," in Proc. Int. Conf. on Big Data, Knowledge and Control Systems Engineering, 2025, pp. 1-6.

[2.7] A. Kosar, W. Daelemans, and G. De Pauw, Dont Make Me Guess: Automatically Detecting and Naming Topics in Large Collections of Text. Antwerp, Belgium: Univ. of Antwerp, Faculty of Arts, 2025.

[2.8] J. Van Nooten and W. Daelemans, The Many Faces of a Text: Applications and Enhancements of Multi-Label Text Classification Algorithms. Antwerp, Belgium: Univ. of Antwerp, Faculty of Arts, 2025.

[2.9] T. Bosi, Design, Implementation and Benchmarking of a Retrieval-Augmented Chatbot for the Insurance Sector, Master's thesis, Univ. of Bologna, 2025.

[2.10] J. Karkoush and M. Ali, Källgranskning med RAG och språkmodeller, Student thesis (Basic level, 15 HE credits), Univ. of Gävle, 2025.

[3] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, „Multi-Agent Coordination Strategies vs Retrieval-Augmented Generation in LLMs: A Comparative Evaluation," Electronics, vol. 14, no. 24, 2025, doi:10.3390/electronics14244883.

Цитирана от:

[3.1] M. Nababan and G. Simarmata, "Model Matematika Dalam Pemilihan Mekanisme Koordinasi...," Jurnal Ilmiah Matematika (JIMAT), vol. 6, no. 2, pp. 891-902, Dec. 2025, doi:10.63976/jimat.v6i2.1201.

[3.2] S. Schmulling and G. Sanrocco, "Ensembles of Small Language Models as an Efficient Alternative to Large Language Models," course report (II2202, Fall 2025), KTH Royal Inst. of Technology, Stockholm, Sweden, Jan. 14, 2026.

[4] M. Dimitrova, „Retrieval-Augmented Generation (RAG): Advances and Challenges," PECR, vol. 83, Jul. 2025, doi:10.7546/PECR.83.25.03.

Цитирана от:

[4.1] M. E. Koutsiaki, M. Delianidi, C. Mizeli, K. Diamantaras, I. Grigoropoulos, and N. Koutlianos, "From Textbook to Talkbot: A Case Study of a Greek-Language RAG-Based Chatbot in Higher Education," arXiv:2601.14265, 2025.

BIBLIOGRAPHY

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter,

C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[3] Z. Ji, N. Qiu, S. Xu, D. Young, F. Tao, L. Lyu, C. Chen, C. Gu, R. Li, L. Yang, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[4] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, Nov. 2021, doi: 10.18653/v1/2021.findings-emnlp.320.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[6] N. Rossi, G. M. Gupta, S. Agarwal, S. Srinivasan, J. Liu, S. Han, and Y. Gao, "Relevance filtering for embedding-based retrieval," in *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, 2024, doi: 10.1145/3627673.3680095.

[7] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical domain hallucination test for large language models," in *Proc. 27th Conf. on Computational Natural Language Learning (CoNLL)*, pp. 314–334, 2023, doi: 10.18653/v1/2023.conll-1.21.

[8] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, "ChatLaw: Open-source legal large language model with integrated external knowledge bases," *arXiv preprint arXiv:2306.16092*, Jun. 2023, doi: 10.48550/arXiv.2306.16092.

[9] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, S. Barezi, P. Pascual, H. Li, R. Shick, S. Joty, B. Shin, and P. Fung, "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in *Proc. Int. Joint Conf. on Natural Language Processing and the Asia-Pacific Chapter of the ACL (IJCNLP-AACL)*, 2023.

[10] Y. Gao, Y. Xiong, X. Wang, J. Wang, Z. Jiang, H. Li, Y. Wen, K. Jiang, N. Meng, L. Shao, and P. Sethi, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, Dec. 2023, doi: 10.48550/arXiv.2312.10997.

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[12] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452–454, May 2016, doi: 10.1038/533452a.

[13] R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, and J. Lin, "Initial nugget evaluation results for the TREC 2024 RAG track with the AutoNuggetizer framework," *arXiv preprint arXiv:2411.09607*, Nov. 2024, doi: 10.48550/arXiv.2411.09607.

[14] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, vol. 38, no. 16, pp. 17754–17762, 2024.

[15] I. Radeva, I. Popchev, and M. Dimitrova, "Similarity thresholds in retrieval-augmented generation," in *2024 IEEE 12th Int. Conf. on Intelligent Systems (IS)*, Aug. 2024, pp. 1–7, doi: 10.1109/IS61756.2024.10705214.

[16] M. Dimitrova, I. Popchev, and I. Radeva, "PaSSER: A platform for evaluating LLMs in RAG," in *2025 IEEE BdKCSE*, 2025, p. 7, doi: 10.1109/BdKCSE67969.2025.11300500.

[17] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, "Web application for retrieval-augmented generation: Implementation and testing," *Electronics*, vol. 13, no. 7, p. 1361, Apr. 2024, doi: 10.3390/electronics13071361.

[18] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, "Multi-agent coordination strategies vs. retrieval-augmented generation in LLMs: A comparative evaluation," *Electronics*, vol. 14, no. 24, p. 4883, Jan. 2025, doi: 10.3390/electronics14244883.

[19] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. draft. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/ (accessed Jan. 20, 2026).

[20] M. Dimitrova, "Retrieval-augmented generation (RAG): Advances and challenges," Problems of Engineering Cybernetics and Robotics (PECR), vol. 83, Jul. 2025, doi: 10.7546/PECR.83.25.03.

[21] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.

[22] J. Huang, X. Chen, S. Mishra, H. S. Liao, J. J. Chung, H. G. Song, and D. Zhou, "Large language models cannot self-correct reasoning yet," *arXiv preprint arXiv:2310.01798*, Oct. 2023, doi: 10.48550/arXiv.2310.01798.

[23] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP*, Nov. 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[25] A. Shrivastava and P. Li, "Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS)," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014.

[26] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Sep. 2021, doi: 10.1109/TBDATA.2019.2921572.

[27] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, Apr. 2020, doi: 10.1109/TPAMI.2018.2889473.

[28] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2020, doi: 10.18653/v1/2020.acl-main.703.

[29] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014.

[30] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, J. Polosukhin, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M. Kelcey, M. W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Trans. Assoc. Comput. Linguist.*, vol. 7, pp. 452–466, 2019, doi: 10.1162/tacl_a_00276.

[31] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2017, pp. 1601–1611, doi: 10.18653/v1/P17-1147.

[32] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and VERification," in *Proc. NAACL-HLT*, 2018, pp. 809–819, doi: 10.18653/v1/N18-1074.

[33] P. Bajaj, D. Campos, N. Craswell, L. Deng, C. Majumder, X. Qu, B. de Rossi, A. Rodriguez, B. Bhaskar, R. Lin, S. Sayyaparaju, and J. Shao, "MS MARCO: A human generated MAchine reading COmprehension dataset," *arXiv preprint arXiv:1611.09268*, Nov. 2016, doi: 10.48550/arXiv.1611.09268.

[34] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, Oct. 2023, doi: 10.48550/arXiv.2310.11511.

[35] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," *arXiv preprint arXiv:2401.15884*, Jan. 2024, doi: 10.48550/arXiv.2401.15884.

[36] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity," in *Proc. NAACL-HLT (Long Papers)*, 2024,

pp. 7036–7050.
[37] Z. Jiang, F. F. Xu, L. Gao, J. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, C. Callison-Burch, and G. Neubig, "Active retrieval augmented generation," in *Proc. EMNLP*, 2023, pp. 7969–7992, doi: 10.18653/v1/2023.emnlp-main.495.
[38] "Reducing false positives in retrieval-augmented generation (RAG) semantic caching: A banking case study," *InfoQ*. Accessed: Jan. 21, 2026. [Online]. Available: https://www.infoq.com/articles/reducing-false-positives-retrieval-augmented-generation/
[39] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," *arXiv preprint arXiv:2006.14799*, Jun. 2020, doi: 10.48550/arXiv.2006.14799.
[40] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proc. EACL (Demos)*, 2024.
[41] R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, and J. Lin, "Ragnarök: A reusable RAG framework and baselines for TREC 2024 retrieval-augmented generation track," *arXiv preprint arXiv:2406.16828*, Jun. 2024, doi: 10.48550/arXiv.2406.16828.
[42] "Getting Started," *TruLens*. Accessed: Jan. 31, 2026. [Online]. Available: https://www.trulens.org/getting_started/
[43] "LangSmith Evaluation," *LangChain Docs*. Accessed: Jan. 31, 2026. [Online]. Available: https://docs.langchain.com/langsmith/evaluation
[44] "Observability concepts," *LangChain Docs*. Accessed: Jan. 31, 2026. [Online]. Available: https://docs.langchain.com/langsmith/observability-concepts
[45] "Home," *Arize Phoenix*. Accessed: Jan. 31, 2026. [Online]. Available: https://phoenix.arize.com/
[46] "DeepEval," *Confident AI*. Accessed: Jan. 31, 2026. [Online]. Available: https://deepeval.com/
[47] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," *arXiv preprint arXiv:2401.05856*, Jan. 2024, doi: 10.48550/arXiv.2401.05856.
[48] T. Yu, S. Zhang, and Y. Feng, "Auto-RAG: Autonomous retrieval-augmented generation for large language models," *arXiv preprint arXiv:2411.19443*, Nov. 2024, doi: 10.48550/arXiv.2411.19443.
[49] D. Edge, H. Trinh, B. Cheng, J. Bradley, N. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph RAG approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, Apr. 2024, doi: 10.48550/arXiv.2404.16130.
[50] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, "LightRAG: Simple and fast retrieval-augmented generation," *arXiv preprint arXiv:2410.05779*, Oct. 2024, doi: 10.48550/arXiv.2410.05779.
[51] Z. Wang, J. Cho, S. S. Kim, S. J. Hwang, S. Lee, and J. G. Park, "Speculative RAG: Enhancing retrieval augmented generation through drafting," *arXiv preprint arXiv:2407.08223*, Jul. 2024, doi: 10.48550/arXiv.2407.08223.
[52] I. Popchev, L. Doukovska, and I. Radeva, "A prototype of blockchain/distributed file system platform," in *2022 IEEE 11th Int. Conf. Intell. Syst. (IS)*, Oct. 2022, pp. 1–7, doi: 10.1109/IS57118.2022.10019715.
[53] I. Popchev, L. Doukovska, and I. Radeva, "A framework of blockchain/IPFS-based platform for smart crop production," in *2022 Int. Conf. Automatics and Informatics (ICAI)*, Oct. 2022, pp. 265–270, doi: 10.1109/ICAI55857.2022.9960070.
[54] I. Popchev and I. Radeva, "Decentralized application (dApp) development and implementation," *Cybernetics and Information Technologies*, vol. 24, no. 2, pp. 122–141, Jun. 2024, doi: 10.2478/cait-2024-0019.
[55] AntelopeIO, "Antelope," GitHub repository. Accessed: Jun. 14, 2025. [Online]. Available: https://github.com/AntelopeIO
[56] IPFS, "IPFS Documentation," Accessed: Jun. 14, 2025. [Online]. Available: https://docs.ipfs.tech/

[57] I. Popchev, I. Radeva, and L. Doukovska, "Oracles integration in blockchain-based platform for smart crop production data exchange," *Electronics*, vol. 12, no. 10, Art. no. 2244, Jan. 2023, doi: 10.3390/electronics12102244.
[58] Greymass, "greymass/anchor: Antelope Desktop Wallet and Authenticator," GitHub repository. Accessed: Jan. 08, 2026. [Online]. Available: https://github.com/greymass/anchor
[59] EOSio Support, "Anchor Wallet Overview," Accessed: Mar. 15, 2023. [Online]. Available: https://eosio.support/anchor-wallet-overview/
[60] Chroma, "Chroma," Accessed: Jan. 21, 2026. [Online]. Available: https://www.trychroma.com
[61] Chroma Research, "Evaluating chunking strategies for retrieval," Accessed: Feb. 04, 2026. [Online]. Available: https://research.trychroma.com/evaluating-chunking
[62] LangChain, "LangChain," Accessed: Jun. 14, 2025. [Online]. Available: https://www.langchain.com
[63] Ollama, "Ollama," Accessed: Jan. 21, 2026. [Online]. Available: https://ollama.com
[64] PyPI, "pyntelope," Accessed: Jun. 14, 2025. [Online]. Available: https://pypi.org/project/pyntelope/
[65] Modal, "How much VRAM do I need for LLM inference?," Accessed: Feb. 04, 2026. [Online]. Available: https://modal.com/blog/how-much-vram-need-inference
[66] J. Manchanda, L. Boettcher, M. Westphalen, and J. Jasser, "The open source advantage in large language models (LLMs)," *arXiv preprint arXiv:2412.12004*, Dec. 2024, doi: 10.48550/arXiv.2412.12004.
[67] AI21, "What is a long context window? Benefits & use cases," Accessed: Feb. 04, 2026. [Online]. Available: https://www.ai21.com/knowledge/long-context-window/
[68] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.
[69] Mistral AI, "Announcing Mistral 7B," Accessed: Jan. 21, 2026. [Online]. Available: https://mistral.ai/news/announcing-mistral-7b/
[70] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Itani, K. Jiotomo, M. Levy, W. Luo, T. Scialom, G. Sun, K. S. Balaji, A. Sagun, E. Grave, S. Goyal, T. Izacard, A. Kushman, P. Luc, S. Iyer, A. Lomeli, Y. Low, J. Martin, P. Bhargava, M. Sastry, S. Singh, M. Singh, T. Majid, R. Williams, T. Scialom, and J. Zettlemoyer, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, Jul. 2023, doi: 10.48550/arXiv.2307.09288.
[71] A. Mitra, H. S. Liao, M. Moussawi, A. S. Atanasova, A. S. Sestari, H. Song, J. G. Park, J. J. Chung, and J. Huang, "Orca 2: Teaching small language models how to reason," *arXiv preprint arXiv:2311.11045*, Nov. 2023, doi: 10.48550/arXiv.2311.11045.
[72] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Castro, M. S. Lauw, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, Oct. 2023, doi: 10.48550/arXiv.2310.06825.
[73] R. Rastogi, "Papers explained: Mistral 7B," DAIR.AI (Medium). Accessed: Mar. 06, 2024. [Online]. Available: https://medium.com/dair-ai/papers-explained-mistral-7b-b9632dedf580
[74] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
[75] GSM8K, "openai/grade-school-math," GitHub repository. Accessed: Feb. 04, 2026. [Online]. Available: https://github.com/openai/grade-school-math
[76] M. Suzgun, N. S. Abid, A. Adam, E. Ahumada, A. Bansal, T. B. Brown, W. J. Child, E. Choi, D. S. Weld, and L. Zettlemoyer, "Challenging BIG-Bench tasks and whether chain-of-thought can solve them," *arXiv preprint*

arXiv:2210.09261, Oct. 2022, doi: 10.48550/arXiv.2210.09261.
[77] IBM, "IBM Granite 3.2: open source reasoning and vision," Accessed: Jan. 21, 2026. [Online]. Available: https://www.ibm.com/new/announcements/ibm-granite-3-2-open-source-reasoning-and-vision
[78] DeepSeek-AI, "deepseek-ai/DeepSeek-R1," GitHub repository. Accessed: Jan. 21, 2026. [Online]. Available: https://github.com/deepseek-ai/DeepSeek-R1
[79] Meta AI, "Introducing Llama 3.1: Our most capable models to date," Accessed: Jan. 21, 2026. [Online]. Available: https://ai.meta.com/blog/meta-llama-3-1/
[80] Mistral AI, "Mistral 7B," Accessed: Jan. 21, 2026. [Online]. Available: https://docs.mistral.ai/models/mistral-7b-0-3
[81] DeepSeek-AI, C. Guo, M. Yang, Z. Bi, K. Zhou, F. Wang, W. Liu, Z. Shao, D. Wang, and G. Dai, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, Jan. 2025, doi: 10.48550/arXiv.2501.12948.
[82] A. Grattafiori, J. Santua, K. Stone, P. Albert, S. Batra, K. J. Chen, A. Chou, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Itani, K. Jiotomo, A. Kushman, P. Luc, M. Martin, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, B. Fuller, C. Gao, V. Goswami, and N. Goyal, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, Jul. 2024, doi: 10.48550/arXiv.2407.21783.
[83] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[84] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, Jun. 2005, pp. 65–72.
[85] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81.
[86] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318, doi: 10.3115/1073083.1073135.
[87] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *Intell. Data Eng. Autom. Learn. (IDEAL 2013)*, *Lecture Notes in Computer Science*, vol. 8206, pp. 611–618, 2013, doi: 10.1007/978-3-642-41278-3_74.
[88] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
[89] H. Kane, M. Y. Kocyigit, A. Abdalla, P. Ajanoh, and M. Coulibali, "NUBIA: NeUral based interchangeability assessor for text generation," in *Proc. 1st Workshop on Evaluating NLG Evaluation*, Dec. 2020, pp. 28–37.
[90] T. Ito, K. van Deemter, and J. Suzuki, "Reference-free evaluation metrics for text generation: A survey," *arXiv preprint arXiv:2501.12011*, Jan. 2025, doi: 10.48550/arXiv.2501.12011.
[91] D. C. Montgomery, *Statistical Quality Control: A Modern Introduction*, 6th ed. Hoboken, NJ, USA: Wiley, 2010.
[92] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY, USA: Psychology Press, 2009.
[93] Regulation (EU) 2018/848 of the European Parliament and of the Council of 30 May 2018 on organic production and labelling of organic products and repealing Council Regulation (EC) No 834/2007. Accessed: Jan. 21, 2026. [Online]. Available: http://data.europa.eu/eli/reg/2018/848/oj
[94] FAO, "Climate Smart Agriculture Sourcebook," Accessed: Jan. 21, 2026. [Online]. Available: https://www.fao.org/climate-smart-agriculture-sourcebook/en/
[95] Convention on Biological Diversity, "The Convention on Biological Diversity," Accessed: Jan. 07, 2026.

[Online]. Available: https://www.cbd.int/convention
[96] European Commission, "Biodiversity Strategy for 2030," Accessed: Jan. 07, 2026. [Online]. Available: https://environment.ec.europa.eu/strategy/biodiversity-strategy-2030_en