МРНТИ 50.41.29 **https://doi.org/10.58805/kazutb.v.4.29-1029**

СРАВНЕНИЕ ПОДХОДОВ ДОПОЛНЕННОЙ ПОИСКОМ ГЕНЕРАЦИИ ДЛЯ СИСТЕМ АНАЛИЗА НОРМАТИВНО-ПРАВОВЫХ АКТОВ КАЗАХСТАНА

И.Н. Букенова, С.И. Компан, Г.С. Букенов, Б.А. Казангапова

Алматинский технологический университет, Алматы, Казахстан

Корреспондент-автор: ibukenowa@mail.ru

Фундаментальными ограничениями для LLM остаются их статичные знания и, следовательно, риск генерации недостоверной информации, “галлюцинаций”. Технология Retrieval-Augmented Generation (RAG) была предложена для преодоления этих проблем путем интеграции с внешними базами знаний с проверенной актуальной информацией. Однако стандартные RAG парадигмы на основе векторов оказались неэффективны для сложных запросов, в которых необходимы более глубокие размышления или синтез информации из нескольких фрагментов. Как перспективная альтернатива рассматривается GraphRAG, вид RAG использующий структурированное представление знаний в виде графов. Но вопрос сравнительной эффективности, баланса качества и ресурсов для GraphRAG против векторных систем, особенно в специфических направлениях (как правовая система Казахстана), исследован еще недостаточно. Данная работа представляет количественный эксперимент: мы сравнили базовый векторный RAG с двумя конфигурациями Microsoft GraphRAG (локальным/глобальным поиском), используя gpt-4.1-nano и text-embedding-3-small для создания индексов. Эксперимент проводился на корпусе 15 НПА РК (10.1 млн. символов, 1.24 млн. слов, ~28,700 фрагментов). Был создан синтетический датасет из 250 вопросов - 50 одношаговых, 100 многошаговых, 100 тематических. Для оценки применялся комплекс метрик: качество поиска (NDCG@10, Context Recall/Precision), качество ответа (LLM-as-Judge оценка по релевантности, точности, полноте, обоснованности) и ресурсы (стоимость/задержка). Результаты ($p<0.05$) свидетельствуют: GraphRAG с локальным поиском (оптимально при доле извлечения текста 0.50) статистически превосходит векторный RAG по качеству извлечения в multi-hop задачах (NDCG@10: 0.657 vs 0.555, +18.6%; Context Recall: 0.710 vs 0.544, +16.6%, при пиковой Context Precision 0.784). GraphRAG с глобальным поиском лидирует по качеству итоговых ответов для тематических вопросов (Quality Score: 3.89; +22.3% vs векторного RAG). Анализ параметров локального поиска GraphRAG подтвердил наличие явного Парето-фронта “качество-ресурсы”. На основе полученных результатов мы сделали выводы: графовые системы формируют границу производительности для сложных запросов, но ценой больших затрат, векторный RAG сравнительно хорошо справляется с одиночными вопросами, при этом экономичнее. Исследование дает эмпирические основания для выбора RAG-архитектуры в зависимости от предпочтений в юридической сфере, учитывая тип запроса и ресурсы.

Ключевые слова: генерация дополненная поиском, информационный поиск, большие языковые модели, графовые представления знаний, векторные представления знаний

COMPARISON OF SEARCH-AUGMENTED GENERATION APPROACHES FOR REGULATORY ANALYSIS SYSTEMS IN KAZAKHSTAN

I.N. Bukenova, S.I. Kompan, G.S. Bukenov, B.A. Kazangapova

Almaty Technological University, Almaty, Kazakhstan,
e-mail: ibukenowa@mail.ru

The main technical challenges that affect LLMs include frozen knowledge maintenance and the generation of “hallucinations.” RAG tries to tackle the problem through external knowledge base linkage to models which verify their information sources. Our investigation exposed that standard vector-based

RAG struggles when handling combined information retrieval tasks from multiple knowledge sources. The knowledge organization system GraphRAG presents another option instead of standard vector-based systems. Research needs more extensive evaluation of GraphRAG performance relative to vector-based methods and their quality versus cost impact in Kazakhstan's legal context. An experiment was conducted to assess Microsoft GraphRAG (local and global search configurations) when compared to a baseline vector RAG for text generation purposes with gpt-4.1-nano and text-embedding-3-small indexing model selection. A total of 15 Kazakhstani legal documents served as the corpus basis containing 10.1M characters together with 1,24M words and 28,700 fragments. After creating 250 test questions we separated them into single-step (50) and multi-step (100) and thematic (100) categories. The assessment analyzed how well information was retrieved using NDCG@10 and Context Recall/Precision measurements as well as how well the system performed in generating answers (relevance, accuracy, completeness, and justification) and the allocation of resources against evaluation tasks (cost, latency). Significant findings emerged ($p<0.05$). The best performance was achieved by GraphRAG with local search when operating at 50% text extraction because it provided superior results to vector RAG throughout multi-step queries (+18.6% NDCG@10 to 0.657; +16.6% Context Recall at 0.710; peak Context Precision 0.784). The global search GraphRAG variant produced answers with superior quality ratings for thematic questions (3.89 Quality Score which represented a 22.3% better score than vector RAG). Science-based optimization proved that GraphRAG creates a separate area in the quality-resource design space. The increased complexity level of using graph-based systems comes with associated higher implementation expenses. Vector RAG stands as an economically advantageous solution for basic information queries. The obtained results help engineers pick RAG architectures for legal applications through combinations of query requirements and system resource guidelines.

Keywords: augmented Search Generation, Information Search, Large Language Models, Graph Representations of Knowledge, Vector representations of Knowledge

ҚАЗАҚСТАННЫҢ НОРМАТИВТІК-ҚҰҚЫҚТЫҚ АКТІЛЕРІН ТАЛДАУ ЖҮЙЕЛЕРІ ҮШІН ІЗДЕУМЕН ТОЛЫҚТЫРЫЛҒАН ГЕНЕРАЦИЯ ТӘСІЛДЕРІН САЛЫСТЫРУ

И.Н. Букенова✉, С.И. Компан, Г.С. Букенов, Б.А. Казангапова

Алматы технологиялық университеті, Алматы, Қазақстан,
e-mail: ibukenowa@mail.ru

LLM үшін іргелі шектеулер олардың статикалық білімі болып қала береді, сондықтан жалған ақпарат, "галлюцинация"пайда болу қаупі бар. Retrieval-Augmented Generation (RAG) технологиясы дәлелденген өзекті ақпаратпен сыртқы білім базаларымен біріктіру арқылы осы қиындықтарды жеңу үшін ұсынылды. Дегенмен, векторларға негізделген стандартты Rag парадигмалары тереңірек ойлауды немесе бірнеше үзінділерден ақпаратты синтездеуді қажет ететін күрделі сұраулар үшін тиімсіз болып шықты. GraphRAG перспективті альтернатива ретінде білімнің графиктер түріндегі құрылымдық көрінісін пайдаланатын RAG түрі болып саналады.Бірақ салыстырмалы тиімділік, сапа балансы және graphrag үшін ресурстар мәселесі векторлық жүйелерге қарсы, әсіресе нақты бағыттарда (Қазақстанның құқықтық жүйесі ретінде) әлі зерттелмеген. Бұл жұмыс сандық экспериментті ұсынады: біз индекстерді құру үшін GPT-4.1-nano және text-embedding-3-small көмегімен негізгі векторлық RAG-ді екі Microsoft GraphRAG конфигурациясымен (жергілікті/Ғаламдық іздеу) салыстырдық. Эксперимент ҚР 15 НҚА корпусында жүргізілді (10.1 млн.таңба, 1.24 млн. сөз, ~28,700 фрагмент).250 сұрақтан тұратын синтетикалық деректер жиынтығы құрылды - 50 бір сатылы, 100 көп сатылы, 100 тақырыптық. Бағалау үшін көрсеткіштер кешені қолданылды: іздеу сапасы (ndcg@10, Context Recall/Precision), жауап сапасы (LLM-as-Judge өзектілігі, дәлдігі, толықтығы, дұрыстығы бойынша бағалау) және ресурстар (cost/latency). Нәтижелер ($P<0.05$) мынаны көрсетеді: жергілікті іздеумен GraphRAG (0.50 мәтінді шығару үлесі үшін оңтайлы) көп саты-

лы есептердегі экстракция сапасы бойынша векторлық RAG-дан статистикалық түрде асып түседі (ndcg@10: 0.657 vs 0.555, +18.6%; Context Recall: 0.710 vs 0.544, +16.6%, шыңы кезінде context precision 0.784). Жаһандық іздеуі бар GraphRAG тақырыптық сұрақтарға қорытынды жауаптардың сапасы бойынша көш бастап тұр (Quality Score: 3.89; +22.3% vs векторлық RAG). GraphRAG жергілікті іздеу параметрлерін талдау нақты ”сапа-ресурстар” Парето майданының болуын растады. Алынған нәтижелерге сүйене отырып, біз қорытынды жасадық: графикалық жүйелер күрделі сұраулар үшін өнімділік шекарасын қалыптастырады, бірақ үлкен шығындармен векторлық RAG бір сұрақтармен салыстырмалы түрде жақсы жұмыс істейді, ал үнемді. Зерттеу сұрау түрі мен ресурстарды ескере отырып, заң саласындағы қалауларға негізделген RAG архитектурасын таңдауға эмпирикалық негіз береді.

Түйін сөздер: іздеумен толықтырылған генерация, ақпараттық іздеу, үлкен тілдік модельдер, білімнің графикалық көріністері, білімнің векторлық көріністері

Введение. Генерация, дополненная поиском (RAG) расширяет функционал больших языковых моделей, сочетает их параметрическую память с внешним корпусом знаниям через ретривер [1]. Появление этого подхода обусловлено как раз ограничениями самих моделей: статичностью знаний [2, 3] и склонностью к “галлюцинациям” [4]. Проблема особенно остра в сферах с высоким требованием к точности информации, например в праве, но даже коммерческие решения тут показывают уровень галлюцинаций 17-33% [5]. Извлекая релевантную информацию во время генерации [1, 6], этот подход повышает точность ответов, превосходя, к примеру, BART на 35.6% по фактологичности (42.7% vs 7.1%), на 20.6% по специфичности (37.4% vs 16.8%) на некоторых задачах [1]. Также доминирует над моделями типа T5-11B на наукоемких задачах (+7.9% EM на NQ) [1].

Актуальность исследования обусловлена несколькими факторами. Во-первых, юридическая сфера предъявляет повышенные требования к достоверности информации: даже коммерческие LLM-решения демонстрируют уровень галлюцинаций 17-33% на юридических задачах [5], что недопустимо для практического применения. Во-вторых, правовая система Казахстана характеризуется сложной структурой взаимосвязанных нормативных актов, содержащих многочисленные отсылки между документами, что создает объективную потребность в методах, способных учитывать эти связи при информационном поиске. В-третьих, несмотря на активное развитие технологий подхода генерации дополненной поиском на основе графов, отсутствуют эмпирические данные об их эффективности на корпусах НПА постсоветских юрисдикций, что затрудняет принятие обоснованных решений при разработке систем правовой аналитики для региона СНГ. Наконец, высокая ресурсоемкость графовых методов требует количественной оценки компромиссов между качеством и затратами для обоснования их внедрения в практику.

Практическая значимость работы заключается в предоставлении эмпирических оснований для выбора архитектуры таких-систем при разработке инструментов правовой аналитики, с учетом специфики запросов и доступных вычислительных ресурсов.

Парадигмы подходов развиваются от ранних наивных и продвинутых к модульным подходам, использующих графы для представления и поиска [7], включая основанные на знаниях, на индексах и гибридные типы [2, 3], которые лучше справляются с многошаговыми рассуждениями [2]. Графовый подход выделяет этапы: построение графа, индексация, конфигурация операторов и извлечение/генерация [4]. Индексация начинается с чанкинга (размер/перекрытие влияют на стоимость/полноту [4, 8]) с использованием разных стратегий (от фиксированного размера до RCTS [9] или семантического [10]). На этапе построения графа LLM [8, 11] или OpenIE [12] извлекают сущности, отношения и утверждения. RAKG-подход улуч-

шает графовое конструирование [10]. Используются различные графы: граф фрагментов, деревья (RAPTOR [13]), граф знаний, контекстуальный граф знаний, обогащенный граф знаний [4]. Иерархическая структуризация (например, алгоритм Leiden в Microsoft GraphRAG [8], LLM-кластеризация в ArchRAG [14], деревья кластеров RAPTOR [13], гетерогенные графы NodeRAG [11]) организует граф для высокоуровневого извлечения. Индексация создает эмбеддинги (text-embedding-3 [9], BGE-M3 [10]) для чанков, узлов и резюме сообществ [8, 11, 14], сохраняемые в векторных БД (FAISS HNSW [1], LanceDB, C-HNSW ArchRAG [14]). Для снижения стоимости предложены multi-granular индексы (KET-RAG [15]). Цикл поиска может включать декомпозицию (PlanRAG [12]), расширение (RoK [16]) или извлечение сущностей (NodeRAG [11]). Основной этап - извлечение - выполняется разными методами: стандартный векторный поиск [1], графовые локальный и глобальный поиск [8, 17], обход соседей (K-hop, PPR [11, 12]), GNN-based (GFM-RAG [18]), иерархический (ArchRAG [14], RAPTOR [13]) и гибридный (TREX [13]). Эмпирические данные показывают преимущества локального поиска для multi-hop запросов (+0.72% F1 HotpotQA [17]) и глобального поиска для тематических запросов [8]. Продвинутые методы (GFM-RAG, NodeRAG, ArchRAG, TCR-QF [19], TREX [13]) демонстрируют значительные улучшения (+16.8% R@2, +4.58% Acc, +10% Acc, +29.1% EM соответственно). Пост-обработка включает реранжирование (REBEL улучшает качество ответа [20], но общие реранкеры могут ухудшать результаты в доменах [9]). Итоговая генерация ответа выполняется LLM [1]; для глобального поиска часто используют MapReduce [4, 8]. Фреймворки вроде PAR RAG интегрируют планирование и верификацию (EM 0.75 HotpotQA [12]).

Несмотря на все это, и векторный, и графовый подход сталкиваются с вызовами. Качество извлечения ключевая проблема; оптимизация лишь по релевантности недостаточна “relevance bottleneck” [20]. Уязвимость к проблеме ошибочного извлечения (падение точности до 45.8% [21]) и низкая точность извлечения в сфере права (Precision@1 < 15% [9]) сохраняется. Автоматически построенные графы знаний страдают от потери информации и неполноты [19, 22]. Это снижает производительность (до -14.7% точности при разрыве пути [22]). Качество графа напрямую влияет на результат [10]. Стоимость и эффективность - важный аспект. Построение и индексация графов затратны [13, 15] (графовый ~2.8x дороже векторный [23]). Сложные пайплайны (PAR RAG, глобальный поиск) характеризуются высокой задержкой и стоимостью (иногда >100K токенов/запрос [4, 12]). Оптимизация гиперпараметров сложна из-за множества целей и стоимости оценки [23]. Оценка - нерешенная проблема, стандартные метрики слабо коррелируют с человеческими суждениями (ρ Спирмена ~0.48 [24]), использование языковых моделей в качестве судей ненадежны ($\alpha < 0.16$ [24]), бенчмарки не измеряют сложность [25], оценка OLAP-запросов открыта [13]. LLM трудно синтезировать информацию из множества документов (эффект «потери в середине» [13]), а интеграция рассуждений чревата избыточным обдумыванием [6].

Следует отметить, что большинство существующих исследований систем для юридической сферы сфокусированы на англоязычных корпусах и юрисдикциях общего права [9], тогда как специфика правовых систем СНГ, основанных на романо-германской традиции и характеризующихся иной структурой кодификации, остается недостаточно изученной. Исследования по применению технологий в правовой сфере в регионе СНГ немногочисленны и преимущественно посвящены отдельным прикладным задачам, без систематической оценки современных архитектур.

Эти ограничения определяют актуальность исследования. Цель работы - провести количественную оценку сравнительной эффективности векторных и графовых подходов применительно к корпусу нормативно-правовых актов Республики Казахстан, с выявлением компромиссов между качеством извлечения информа-

ции и вычислительными ресурсами. Задачи исследования: Построить экспериментальную инфраструктуру для сравнения базового векторного и двух конфигураций Microsoft GraphRAG (локальный и глобальный поиск) на корпусе НПА РК. Оценить статистически значимые различия в качестве извлечения контекста (NDCG@10, Context Recall/Precision) между графовым локальным поиском и векторным на multi-hop запросах. Исследовать влияние параметра доли извлечения текста (p(txt)) в локальном поиске на баланс между качеством извлечения и потреблением ресурсов. Установить сравнительную эффективность глобального поиска относительно векторного для тематических запросов по метрикам качества финального ответа. Определить позиционирование исследуемых систем в пространстве «качество-ресурсы» и выявить наличие доминирующего Парето-фронта. Гипотеза исследования: Графовые подходы обеспечивают статистически значимое превосходство над векторными методами в задачах, требующих многошагового рассуждения и тематического обобщения, формируя доминирующий фронт в пространстве «качество-ресурсы», однако ценой существенно больших вычислительных затрат. Научная новизна исследования заключается в следующем: Впервые проведена количественная оценка эффективности графовых подходов применительно к корпусу НПА постсоветской юрисдикции (Республика Казахстан). Установлены статистически значимые закономерности влияния параметра доли извлечения текста p(txt) на баланс качества и ресурсов в локальном поиске, включая выявление нелинейных зависимостей и точки оптимума. Проведен комплексный Парето-анализ позиционирования векторных и графовых систем в многомерном пространстве метрик качества и ресурсов для различных типов запросов. Получены эмпирические свидетельства формирования графовыми системами доминирующего Парето-фронта для сложных multi-hop и тематических запросов в правовой предметной области. Исследовательские вопросы: RQ1. Обеспечивает ли GraphRAG с локальным поиском статистически значимое ($p<0.05$) превосходство над векторным на multi-hop запросах? RQ2. Как изменение параметра доли извлечения текста p(txt) в локальном поиске влияет на метрики качества и ресурсы (latency (задержка отклика), стоимость) при обработке multi-hop запросов? RQ3. Превосходит ли GraphRAG с глобальным поиском векторный по метрикам качества итогового ответа и полноты контекста на тематических запросах ($p<0.05$)? RQ4. Какова конфигурация исследуемых систем в пространстве «качество-ресурсы» и формируют ли графовые подходы доминирующий Парето-фронт относительно векторных методов? исследовательские вопросы RQ.1. Превосходит ли Graph RAG с локальным поиском векторный для multi-hop запросов ($p<0.05$)? 2. Как p(txt) локального поиска влияет на разные метрики для multi-hop запросов запросах? 3. Лучше ли глобальный поиск графового способа, чем векторный на тематических запросах)? (4. Каково позиционирование систем в пространстве “качество-ресурсы” и формируют ли графовые RAG доминирующий фронт?

Материалы и методы. Корпус составили 15 НПА Казахстана: (Административный, Гражданский (Общая и Особенная) части, Гражданский, Уголовный, Уголовно-процессуальный, Предпринимательский, Трудовой, Земельный, О браке (супружестве) и семье, Налоговый кодекс. Также вошли Конституция РК и законы о противодействии коррупции, о правовых актах и о жилищных отношениях). Общий объем: 10.1 млн. символов, 1.24 млн. слов. Исходные HTML мы взяли с «Әділет», для обработки написали парсер, он выполнял обход DOM дерева документов и находил структурные единицы (статьи, пункты, подпункты) на основе регулярных выражений (r”Статья\s+(\d+)”, r”^(\d+)\.”, r”^(\d+)\)”) и анализе тэгов (< p>, < b>, < i>). Каждому фрагменту присваивался идентификатор (он формировался конкатенацией из кода документа, номера статьи и иерархии пунктов/подпунктов (пример: NK_RK_st681_p8_sp1)), извлекался текст, метаданные и ссылки. Была получена коллекция из 28,718 фрагментов. Пример записи: {”fragment_id”: ”NK_RK_st681_p8_sp1”,

..., ”text”: ”...в соответствии со статьями 330, ...”, ”cross_refs”: [{”link_text”: ”статьями 330”, ”raw_href”: ”/rus/docs/K1700000120#z330”}, ...], ...}.

Из-за отсутствия бенчмарка создан синтетический датасет (250 пар “вопрос-ответ”), что является распространенной практикой [9, 25]. Вопросы генерировались с gpt-4.1-nano, 3 типов: single-hop (ответ в одном фрагменте), multi-hop (требуют синтеза из нескольких связанных фрагментов [2, 6]) и thematic (общая тема, OLAP-подобные [13]). Каждая запись в датасете содержит question_id, question, question_type, ground_truth_ids((список fragment_id) и source_fragments_text (тексты). Примеры записей (вопросов) в датасете: для single-hop: {”question_id”: ”q_074”, ”question”: ”Каковы последствия расторжения инвестиционного контракта, по которому я получал инвестиционную субсидию?”, ”question_type”: ”single-hop”, ”ground_truth_ids”: [”PK_RK_st295_p6-1”], ”source_fragments_text”: [”В случае расторжения инвестиционного контракта юридическое лицо Республики Казахстан возмещает в полном объеме сумму инвестиционной субсидии, выплаченной по инвестиционному контракту.”]}, для thematic: {”question_id”: ”234”, ”question”: ”Каковы сроки и порядок уплаты корпоративного подоходного налога?”, ”question_type”: ”thematic”, ”ground_truth_ids”: [”NK_RK_st306_p1”, ”NK_RK_st306_p2”, ”NK_RK_st306_p3”, ”NK_RK_st305_p2_sp1”], ”source_fragments_text”: [”Если иное не установлено настоящей статьей, налогоплательщик осуществляет уплату по корпоративному подоходному налогу по итогам налогового периода не позднее десяти календарных дней после срока...”, ”Сумма корпоративного подоходного налога, подлежащего уплате в бюджет, уменьшается на сумму корпоративного подоходного налога...”, ”Налогоплательщик, удерживающий корпоративный подоходный налог у источника выплаты...”, ”уплату исчисленных авансовых платежей по корпоративному подоходному налогу…”]}. Для multi-hop: {”question_id”: ”q_049”, ”question”: ”Как происходит погашение налоговой задолженности, если умерший человек был индивидуальным предпринимателем?”, ”question_type”: ”multi-hop”, ”ground_truth_ids”: [”NK_RK_st66_p9”, ”NK_RK_st47_p2”], ”source_fragments_text”: [”Налоговая задолженность индивидуального предпринимателя или лица, занимающегося частной практикой, прекращающих деятельность, погашается за счет денег указанного индивидуального предпринимателя или лица, занимающегося частной практикой, в том числе полученных от реализации его имущества, в порядке очередности, установленной законами Республики Казахстан.”, ”В случае если наследник (наследники) является (являются) несовершеннолетним (несовершеннолетними), то обязательство по погашению налоговой задолженности физического лица, образовавшейся на день его смерти или на дату объявления его умершим, в пределах стоимости наследуемого имущества и пропорционально доле в наследстве на дату его получения возлагается на такого (таких) наследника (наследников) только на основании вступившего в силу решения суда.”]}.

Индекс построен фреймворком Microsoft GraphRAG [8]. Использовались gpt-4.1-nano и модель text-embedding-3-small. Текстовые фрагменты (TextUnits, ~35,000 после чанкинга длинных текстов (chunk_size=1024, chunk_overlap=100)) индексировались. Чанки группировались по fragment_id. Граф строился извлечением сущностей (11 типов: Person, Organization, Location, Event, Statute, Regulation, CaseLaw, Term, Penalty, Right, Obligation) и семантических отношений с помощью LLM [8]. Извлеченные ранее кросс-ссылки были добавлены как структурные ребра. Итоговый граф включал узлы сущностей и текстовые единицы, а также вышеперечисленные типы ребер. Параметры экстракции: max_gleanings=2, batch_size=4. Применялся алгоритм Leiden для иерархической кластеризации графа (L1, L2) [4, 8]. Генерировались эмбеддинги для TextUnits, сущностей, отношений и резюме кластеров [8]. Финальный индекс состоял из набора файлов Apache Parquet (entities.parquet,

relationships.parquet, communities.parquet, reports.parquet, text_units.parquet) и векторного индекса LanceDB для эмбеддингов. Векторный индекс для построен библиотекой FAISS [1] на тех же эмбеддингах и TextUnits, использовался IndexHNSWFlat с косинусной метрикой. Все системы использовали gpt-4.1-nano для генерации ответов. Конфигурации: GR-L(p) - графовый локальный [8, 17], поиск от близких к запросу узлов к соседям; параметр p(txt) (доля извлекаемых текстов из сообществ) тестировался для значений 0.25, 0.50, 0.75. GR-G - графовый глобальный [8, 17], поиск по резюме сообществ верхнего уровня (L2), подходит для тематических запросов [13]. VR - стандартный векторный [1], поиск k=5 ближайших фрагментов. Оценка включала: метрики извлечения (NDCG@10, ContextRecall, ContextPrecision, MRR@10, Recall@5, Precision@5); метрики ресурсов (время ответа, стоимость API вызовов по тарифам gpt-4.1-nano-2025-04-14 вход: $0.10 / выход: $0.40); метрику качества ответа (оценка LLM-судьей gpt-4.1-mini по 5-балльной шкале по релевантности, точности, полноте, обоснованности) (такая оценка соответствует современным практикам) [8, 24, 26].

Обсуждение и результаты. В этом разделе представлены результаты. Рассмотрим таблицу 1. Порядок вопросов для всех систем сверху вниз: многошаговые, одношаговые, тематические. Значения метрик указаны средние. стоимость указана умноженная на 1000. Примечание: в таблице не написаны типы вопросов с целью экономии места.

Таблица 1 - Системы - типы вопросов – метрики

Система	NDCG@10	Recall	Precision	Quality	Latency	Стоимость
GR-G	0.677	0.808	0.622	3.688	22289	0.772
	0.634	0.790	0.702	3.625	17941	0.698
	0.702	0.686	0.601	3.888	23437	0.818
GR-L 0.25	0.627	0.611	0.745	3.571	9817	0.350
	0.715	0.702	0.729	3.555	7627	0.311
	0.640	0.628	0.735	3.475	10224	0.394
GR-L 0.5	0.657	0.710	0.784	3.613	15709	0.549
	0.758	0.783	0.751	3.521	13854	0.489
	0.677	0.709	0.769	3.582	19216	0.608
GR-L 0.75	0.618	0.828	0.494	3.620	22533	0.752
	0.682	0.835	0.560	3.433	22973	0.678
	0.620	0.763	0.611	3.380	28222	0.818
VR	0.555	0.544	0.487	3.305	11318	0.220
	0.684	0.789	0.734	3.656	10168	0.209
	0.549	0.624	0.465	3.179	11164	0.243

RQ B1: Сравнение GR-L(0.50) и векторного на multi-hop запросах выявило преимущество графового подхода. Средний NDCG@10 для GR-L(0.50) составил 0.657 против 0.555 у векторного. Средний ContextRecall также выше: 0.710 против 0.544. Визуальное сравнение представлено на Рисунке 1 (верхние панели).

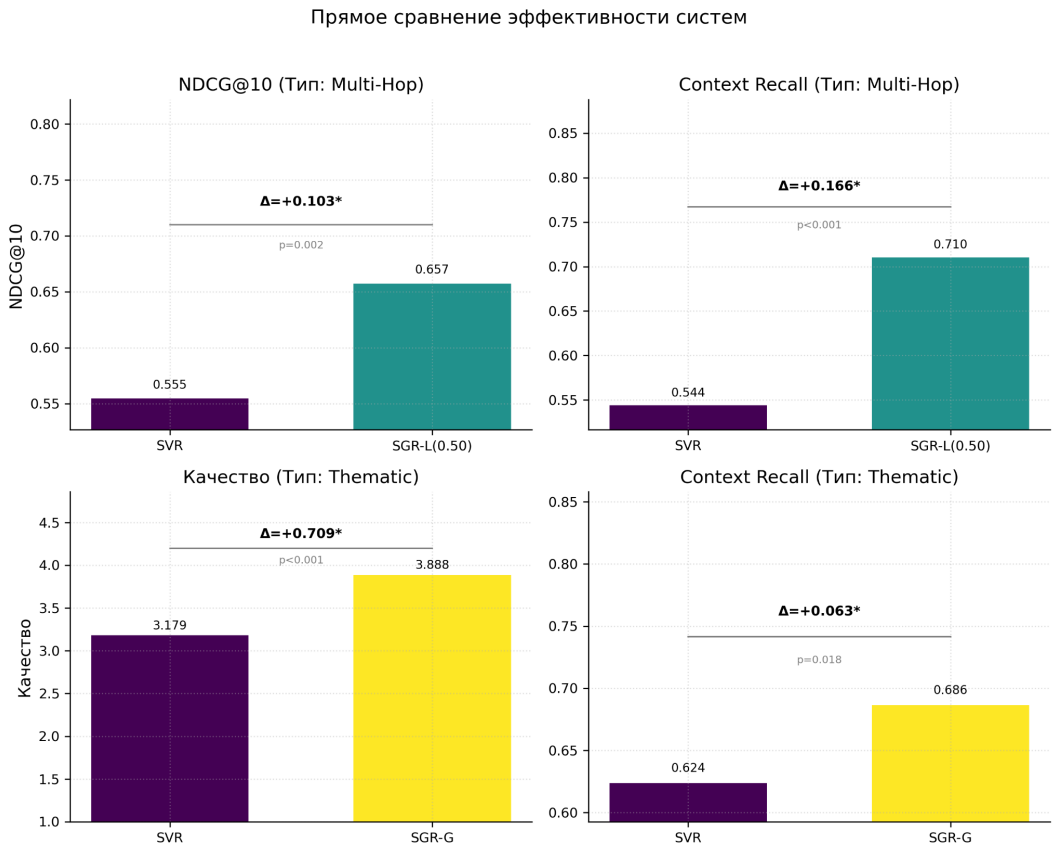


Рис.1 - Сравнение систем на multi-hop запросах (RQ B1, верхние панели) и на тематических запросах (RQ B3, нижние панели)

Таблица 2 - Результаты парного t-теста. Сравнение графового локального с векторным на multi-hop запросах и глобальным графовым с векторным на тематических запросах

Сравнение	Метрика	Статистика t	p-значение	Ср. разница
GR-L 0.25 vs VR	NDCG@10	2.078	0.0202	+0.072
	ContextRecall	2.229	0.0141	+0.067
GR-L 0.50 vs VR	NDCG@10	2.986	0.0018	+0.103
	ContextRecall	5.561	0.0001	+0.166
GR-L 0.75 vs VR	NDCG@10	1.859	0.0330	+0.063
	ContextRecall	10.126	0.0001	+0.284
GR-G vs VR	Quality Score	7.713	0.0001	+0.709
	ContextRecall	2.126	0.0180	+0.063

Статистический анализ подтвердил значимость этих различий (таблица 2). Это свидетельствует о статистически значимом превосходстве графового локального в извлечении релевантного контекста для multi-hop запросов.

Исследование параметра p для графового локального для multi-hop запросов доказало нелинейную зависимость и компромиссы (рисунок 2,

панель A). Средний NDCG@10 достигает максимума (0.657) при p = 0.50, снижаясь при p = 0.75 (0.618). ContextPrecision также максимален при p = 0.50 (0.784) и резко падает при p = 0.75 (0.494). ContextRecall монотонно растет с p (таблица 1). Затраты ресурсов также монотонно увеличиваются с ростом p (таблица 1, Рисунок 3).

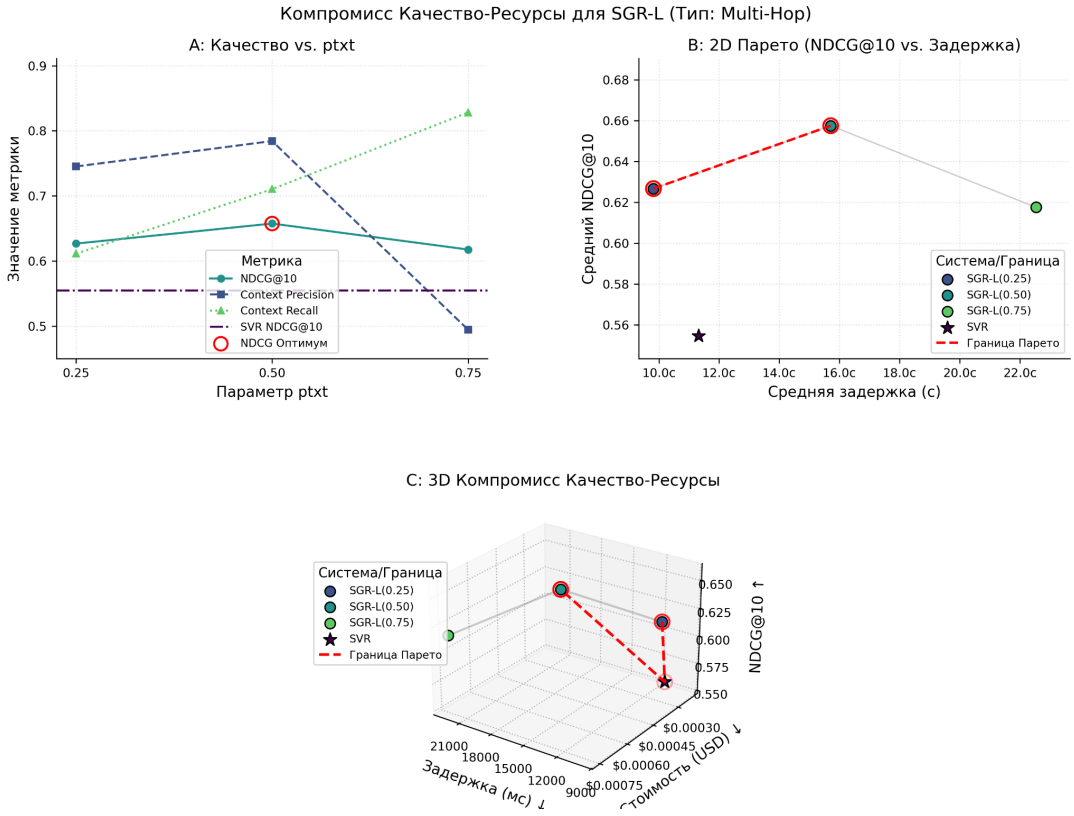


Рис.2 - Анализ компромиссов на multi-hop запросах (RQ B2)

A: Зависимость NDCG@10 и ContextPrecision от p(txt). B: Парето-фронт в пространстве NDCG@10 vs Latency. C: Парето-фронт в пространстве NDCG@10 vs. Задержка vs. Стоимость. Отмечен оптимальный по NDCG@10 вариант графового локального с 0.50

Парето анализ (см. рисунок 2, панели B и C) показывает, что графовые локальные варианты формируют фронт компромисса. Параметр 0.50 представляет собой сбалансированный вариант, наилучшее качество ранжирования NDCG@10 и умеренные затраты ресурсов среди графовых систем с локальным поиском. Векторный находится вне этого фронта, уступая по NDCG@10 при сравнимой или большей задержке, чем с параметром 0.25)

На тематических запросах глобальный графовый показал лучшие результаты по качеству финального ответа. Средняя оценка качества для него составил 3.89 против 3.18 у векторного (Рисунок 1, нижние панели). ContextRecall также был выше у графового глобального (0.686 против 0.624). Статистический анализ (Таблица 2) подтвердил значимость этих различий ($p < 0.0001$ для качества ответа, $p = 0.018$ для ContextRecall). Система глобального обеспечивает более качественные и полные ответы на тематические запросы по сравнению с векторным. Общий анализ производительности систем выявляет четкие профили (Таблица 1, Рисунки 4 - радарные диаграммы, 5, 6). Векторный является наиболее экономичной системой по ресурсам (минимальные средние задержка и стоимость для всех типов запросов). Однако ее показатели качества (NDCG@10, ContextRecall, Ка-

чество) уступают графовым аналогам, особенно на multi-hop и тематических запросах. Локальный демонстрирует гибкость. Изменение параметра позволяет настраивать баланс между качеством и ресурсами.0.25 - наиболее быстрый и дешевый графовый вариант.0.50 часто является оптимальным компромиссом.0.75 обеспечивает максимальную полноту (ContextRecall) ценой снижения точности и увеличения затрат. Глобальный лидирует по качеству ответов и качеству ранжирования (NDCG@10) на тематических запросах. Она также показывает высокие результаты на multi-hop запросах. Однако эта система является наиболее ресурсоемкой (максимальные средние задержка, стоимость).

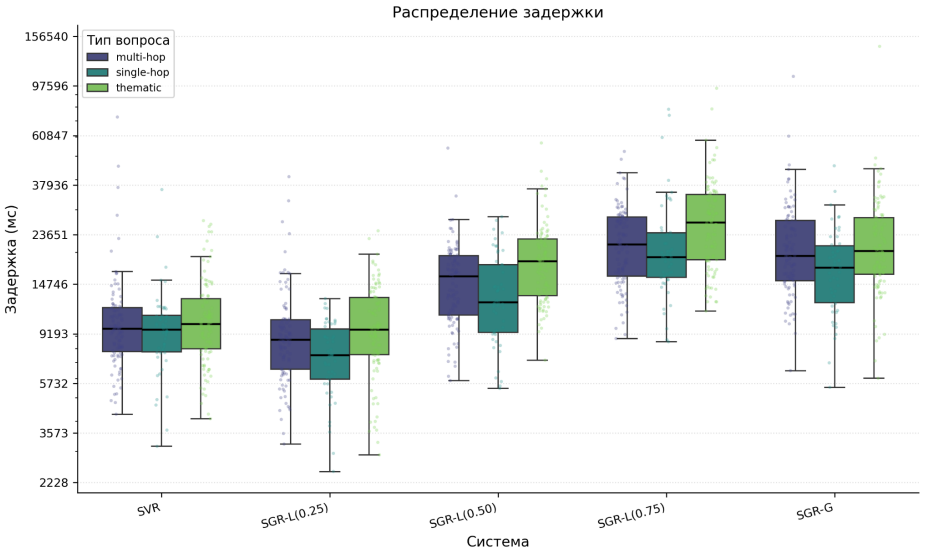


Рис.3 - Распределение времени задержки (ms) по системам и типам запросов (логарифмическая шкала). Показаны медианные значения (k = 1000)

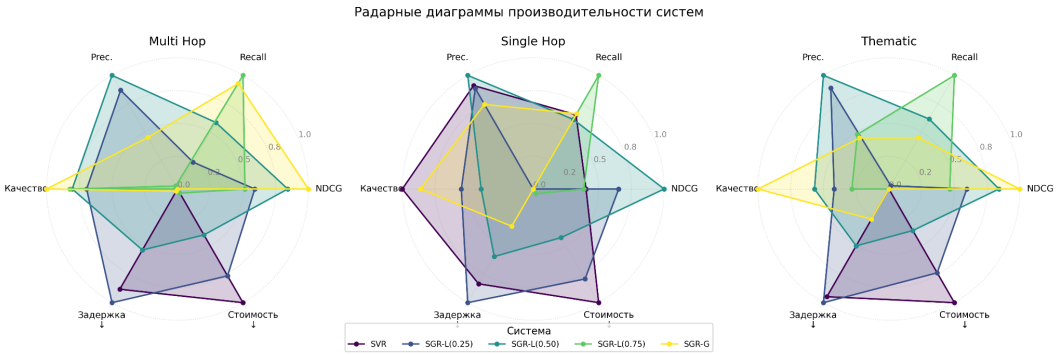


Рис.4 - Радар-диаграммы для сравнения производительности систем по нормализованным метрикам (Precision, Recall, NDCG@10, качество ответа (Quality Score), задержка (Latency), стоимость) для каждого типа вопросов. Оси Задержка и Стоимость инвертированы (лучше - дальше от центра)

Эмпирические кумулятивные функции распределения (Рисунок 6) подтверждают эти наблюдения. Кривые для графовых систем локального и глобального часто смещены вправо (к лучшим значениям) по сравнению с векторным для метрик NDCG@10 и ContextRecall. Для качества доминирует графовые общие.

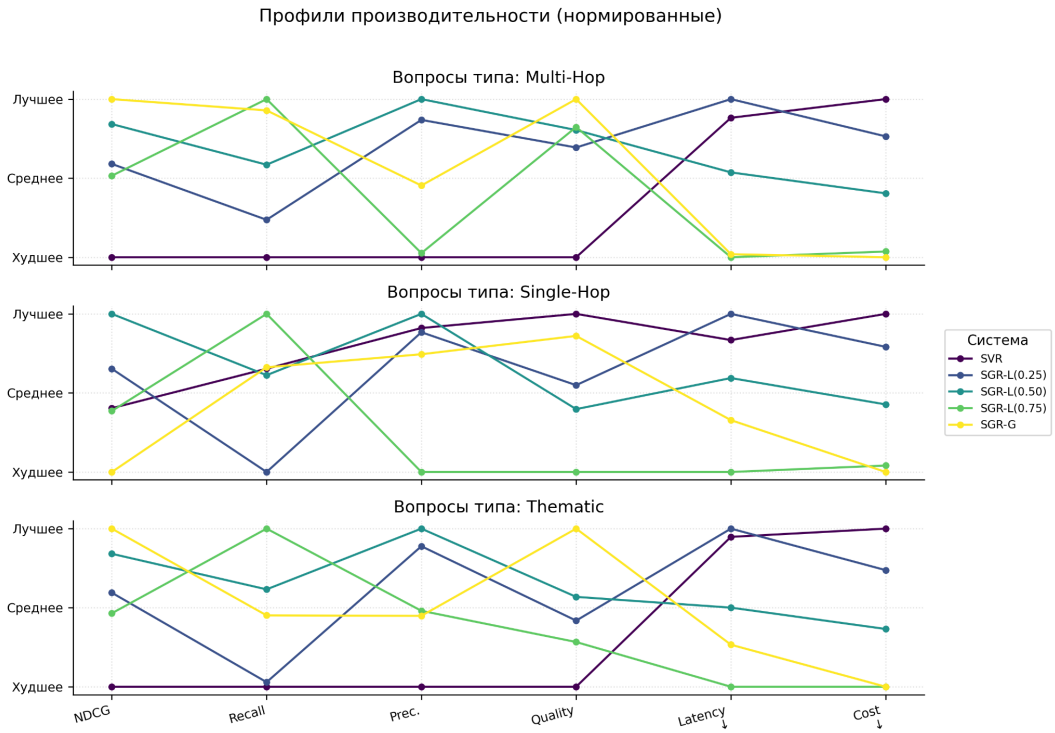


Рис.5 - Диаграммы параллельных координат для визуализации профилей производительности систем по нормализованным метрикам для каждого типа вопросов. Оси ресурсов инвертированы (выше - лучше)

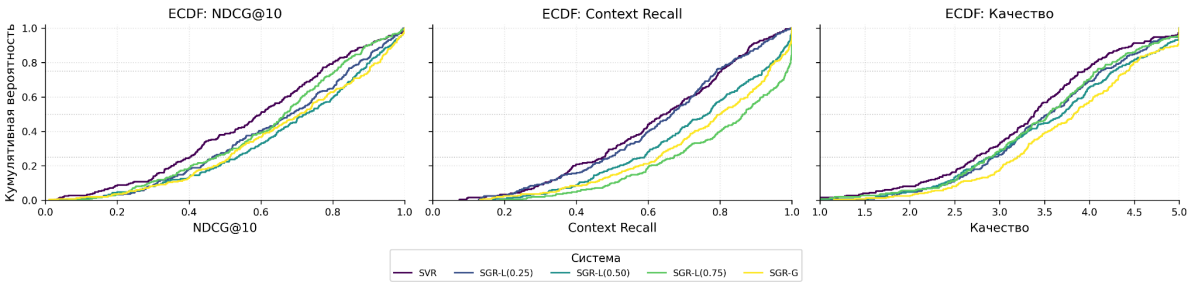


Рис.6 - ECDF для метрик NDCG@10, Context Recall, Качество по типам вопросов

В целом, графовые системы локального, глобального формируют доминирующий Парето-фронт в пространстве ”качество-ресурсы”. Предлагают лучшее качество извлечения по сравнению с векторным, особенно для multi-hop и тематических запросов. Но требуют больших вычислительных ресурсов. Выбор конкретной системы зависит от приоритетов задачи: экономия ресурсов - векторный, сбалансированная производительность - локальный с 0.50 или максимальное качество на тематических запросах - глобальный.

Выводы. Наши результаты эмпирически подтверждают преимущества графовых подходов над векторными для сложных запросов. Локальные статистически достоверно ($p<0.05$) улучшает метрики извлечения NDCG@10 и Recall для multi-hop вопросов по сравнению с векторным. В частности, локальный (с долей извлечения текста 0.50) улучшил NDCG@10 на 0.103 (до 0.657) и Context Recall на 0.166 (до 0.710) относительно векторного.

Это коррелирует с выводами [17] о доминировании локального поиска на multi-hop задачах и подкрепляет гипотезу [2, 3] о пользе явного моделирования связей для multi-hop рассуждений. Возможность локального перемещаться по графу связей, видимо, способствует более эффективному нахождению и агрегации фрагментов для multi-hop запросов, в отличие от векторного, ограниченного поиском ближайших векторных соседей.

Схожим образом, глобальный демонстрирует значимое преимущество над векторным для тематических запросов по оценке Качества (+0.709) и Context Recall (+0.063). Это согласуется с работами [8, 13], характеризующими глобальный поиск как эффективный инструмент для OLAP-запросов или ”sensemaking”, требующих агрегации данных по всему корпусу. Применение агрегированных резюме сообществ [8] в глобальном обеспечивает высокоуровневый контекст для тематических вопросов, чего сложно достичь векторным.

Изучение p для локальных показало компромисс: оптимальное значение 0.50 для NDCG@10 и Context Precision, при монотонном росте Recall (а также Задержки и Стоимости) с увеличением p. Это очерчивает Парето-фронт ”качество-ресурсы”. Это указывает на необходимость оптимизации RAG, что является актуальной темой исследований [23]. Примечательно, что векторная система оказалась доминируемой даже GR-L(0.25) по NDCG@10/Latency, что подчеркивает ограниченную эффективность векторного подхода для комплексных запросов при учете ресурсов. Общий анализ позиционирования систем показал: векторный – самый экономичный, но уступает в качестве на сложных задачах. Глобальный лидирует по качеству тематических ответов, но наиболее ресурсозатратен, что подтверждает данные о стоимости глобального поиска [4].

Локальный предлагает настраиваемый баланс качество/ресурсы. Графовые системы (графовый, глобальный) в целом формируют доминирующий Парето-фронт по качеству для multi-hop и тематических запросов [2]. Наши результаты соответствуют оценке [23], что графовый в среднем требует значительно (~2.8x) больше ресурсов, чем векторный

Практические рекомендации. На основе полученных результатов можно сформулировать следующие рекомендации по выбору архитектуры для задач правовой аналитики. Для систем с ограниченным бюджетом вычислительных ресурсов и преобладанием простых фактологических запросов целесообразно использование базового векторного. Для приложений, требующих обработки сложных multi-hop запросов с синтезом информации из множества связанных фрагментов НПА, рекомендуется GraphRAG с локальным поиском при p(txt)=0.50 как оптимальный компромисс качества и ресурсов. Для аналитических систем, ориентированных на тематические обзоры и обобщения по широким правовым вопросам, оправдано применение GraphRAG с глобальным поиском, несмотря на повышенную ресурсоемкость. При проектировании систем необходимо предусматривать возможность гибридного подхода с выбором метода поиска в зависимости от типа запроса, определяемого автоматически. Настоящее исследование демонстрирует преимущества Microsoft GraphRAG над базовым векторным для корпуса НПА РК. Вместе с тем, следует признать существование более продвинутых графовых архитектур (NodeRAG [11], GFM-RAG [18], RAPTOR [13], PAR RAG [12], ArchRAG [14]), демонстрирующих дальнейшее улучшение различных метрик в англоязычных бенчмарках. Пространство возможных конфигураций систем чрезвычайно обширно, охватывая множество парадигм (наивный, продвинутый, модульный, итеративный, агентный, адаптивный), этапов обработки (индексация, обработка запроса, извлечение, постобработка, генерация) и специфических методов на каждом этапе (например, для обработки запросов: переформулирование, расширение, декомпозиция. Сравнение всех возможных конфигураций выходит за рамки одного исследования.

Этические и практические аспекты применения. Необходимо подчеркнуть, что исполь-

зование таких систем в юридической практике сопряжено с рисками. Несмотря на улучшение показателей точности, графовые и векторные системы не гарантируют полного отсутствия галлюцинаций или ошибочных интерпретаций правовых норм. Применение таких систем должно предусматривать обязательную верификацию результатов квалифицированными юристами, особенно в случаях, влияющих на права и обязанности граждан. Юридическая ответственность за принятие решений на основе рекомендаций LLM-систем остается предметом дискуссии и требует законодательного регулирования.

Разработчики и пользователи подобных систем должны учитывать необходимость обеспечения интерпретируемости результатов и прозрачности процесса принятия решений.

Ограничениями же нашей работы являются: использование корпуса одной юрисдикции; синтетический характер датасета, не отражающий всех реальных сложностей и уязвимостей (например, misleading retrieval [21]); сравнение с ограниченным числом подходов; использование LLM-судьи [13, 24] для оценки качества и отсутствие прямой оценки качества графа [10, 22] или уровня галлюцинаций [5]. Перспективы включают тестирование на реальных запросах, оценку робастности [21, 22], сравнение с более широким спектром архитектур [11, 14, 15, 17, 18], разработку надежных методов оценки [9, 24], изучение методов улучшения графа [10, 19], применение реранжирования [20] и исследование агентных подходов [6, 7] для сложных юридических задач.

Литература

1. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.t., Rocktäschel T., Riedel S. & Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv preprint arXiv:2005.11401.-2020. DOI 10.48550/arXiv.2005.11401.

2. Zhang Q., Chen S., Bei Y., Yuan Z., Zhou H., Hong Z., Dong J., Chen H., Chang Y., Huang X. A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models // arXiv preprint arXiv:2501.13958. -2025. DOI 10.48550/arXiv.2501.13958.

3. Zhu Z., Huang T., Wang K., Ye J., Chen X. & Luo S. Graph-based Approaches and Functionalities in Retrieval-Augmented Generation: A Comprehensive Survey // arXiv preprint arXiv:2504.10499. -2025. DOI 10.48550/arXiv.2504.10499.

4. Zhou Y., Su Y., Sun Y., Wang S., Wang T., He R., Zhang Y., Liang S., Liu X., Ma Y. & Fang, Y. In-depth Analysis of Graph-based RAG in a Unified Framework // arXiv preprint arXiv:2503.04338. -2025. DOI 10.48550/arXiv.2503.04338.

5. Magesh V., Surani F., Dahl M., Suzgun M., Manning C. D. & Ho D. E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools // arXiv preprint arXiv:2405.20362. -2024. DOI 10.48550/arXiv.2405.20362.

6. Gao Y., Xiong Y., Zhong Y., Bi Y., Xue M. & Wang H. Synergizing RAG and Reasoning: A Systematic Review // arXiv preprint arXiv:2504.15909. -2025. DOI 10.48550/arXiv.2504.15909.

7. Singh A., Ehtesham A., Kumar S. & Talaei Khoei T. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG//arXiv preprintarXiv:2501.09136.-2025. DOI 10.48550/arXiv.2501.09136.

8. Edge D., Trinh H., Cheng N., Bradley J., Chao A., Mody A., Truitt S., Metropolitansky D., Ness R. O. & Larson J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization // arXiv preprint arXiv:2404.16130. -2025. DOI 10.48550/arXiv.2404.16130.

9. Pipitone N. & Houir Alami G. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain//arXiv preprint arXiv:2408.10343.-2024. DOI 10.48550/arXiv.2408.10343.

10. Zhang H., Si J., Yan G., Qi B., Cai P., Mao S., Wang D. & Shi B. RAKG: Document-Level Retrieval Augmented Knowledge Graph Construction // arXiv preprint arXiv:2504.09823. -2025. DOI 10.48550/arXiv.2504.09823.

11. Xu T., Zheng H., Li C., Chen H., Liu Y., Chen R. & Sun L. NodeRAG: Structuring Graph-based RAG with Heterogeneous Nodes // arXiv preprint arXiv:2504.11544.-2025. DOI 10.48550/arXiv.2504.11544.

12. Zhang N., Zhang C., Tan Z., Yang X., Deng W. & Wang W. Credible Plan-Driven RAG Method for Multi-Hop Question Answering//arXiv preprint arXiv:2504.16787.-2025. DOI 10.48550/arXiv.2504.16787.

13. Cahoon J., Singh P., Litombe N., Larson J., Trinh H., Zhu Y., Mueller A., Psallidas F., Curino C. Optimizing Open-Domain Question Answering with Graph-Based Retrieval Augmented Generation // arXiv preprint arXiv:2503.02922. -2025. DOI 10.48550/arXiv.2503.02922.

14. Wang S., Fang Y., Zhou Y., Liu X., Ma Y. ArchRAG: Attributed Community-based Hierarchical Retrieval-Augmented Generation//arXiv preprint arXiv:2502.09891.-2025. DOI 10.48550/arXiv.2502.09891.

15. Huang Y., Zhang S. & Xiao X. KET-RAG: A Cost-Efficient Multi-Granular Indexing Framework for Graph-RAG // arXiv preprint arXiv:2502.09304.-2025. DOI 10.48550/arXiv.2502.09304.

16. Yuan X., Liu Y., Di S., Wu S., Zheng L., Meng R., Chen L., Zhou X., Yin J. A Pilot Empirical Study on When and How to Use Knowledge Graphs as Retrieval Augmented Generation // arXiv preprint arXiv:2502.20854. -2025. DOI 10.48550/arXiv.2502.20854.

17. Han H., Shomer H., Wang Y., Lei Y., Guo K., Hua Z., Long B., Liu H., Tang J. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights // arXiv preprint arXiv:2502.11371. -2025. DOI 10.48550/arXiv.2502.11371.

18. Luo L., Zhao Z., Haffari G., Phung D., Gong C., Pan S. GFM-RAG: Graph Foundation Model for Retrieval Augmented Generation // arXiv preprint arXiv:2502.01113. -2025. DOI 10.48550/arXiv.2502.01113.

19. Huang M., Bu C., He Y., Wu X. How to Mitigate Information Loss in Knowledge Graphs for GraphRAG: Leveraging Triple Context Restoration and Query-Driven Feedback//arXiv preprint arXiv:2501.15378. -2025. DOI 10.48550/arXiv.2501.15378.

20. LeVine W., Varjavand B. Relevance Isn’t All You Need: Scaling RAG Systems With Inference-Time Compute via Multi-Criteria Reranking//arXiv preprint arXiv:2504.07104. -2025. DOI 10.48550/arXiv.2504.07104.

21. Zeng L., Gupta R., Motwani D., Yang D., Zhang Y. Worse than Zero-shot? A Fact-Checking Dataset for Evaluating the Robustness of RAG Against Misleading Retrievals//arXiv preprint arXiv:2502.16101. -2025. DOI 10.48550/arXiv.2502.16101.

22. Zhou D., Zhu Y., He Y., Chen J., Kharlamov E., Staab S. Evaluating Knowledge Graph Based Retrieval Augmented Generation Methods Under Knowledge Incompleteness//arXiv preprint arXiv:2504.05163. -2025. DOI 10.48550/arXiv.2504.05163.

23. Barker M., Bell A., Thomas E., Carr J., Andrews T., Bhatt U. Faster, Cheaper, Better: Multi-Objective Hyperparameter Optimization for LLM and RAG Systems//arXiv preprint arXiv:2502.18635. -2025. DOI 10.48550/arXiv.2502.18635.

24. Gienapp L., Hagen T., Fröbe M., Hagen M., Stein B., Potthast M., Scells H. The Viability of Crowdsourcing for RAG Evaluation//arXiv preprint arXiv:2504.15689.-2025. DOI 10.48550/arXiv.2504.15689.

25. Lee J., Kwon D., Jin K., Jeong J., Sim M., Kim M. MHTS: Multi-Hop Tree Structure Framework for Generating Difficulty-Controllable QA Datasets for RAG Evaluation//arXiv preprint arXiv:2504.08756. -2025. DOI 10.48550/arXiv.2504.08756.

26. Gan A., Yu H., Zhang K., Liu Q., Yan W., Huang Z., Tong S. & Hu G. Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey//arXiv preprint arXiv:2504.14891. -2025. DOI 10.48550/arXiv.2504.14891.

Сведения об авторах

Букенова И.Н. - магистр технических наук, лектор, Алматинский технологический университет, Алматы, Казахстан, e-mail: ibukenowa@mail.ru;

Букенов Г.С. - магистр технических наук, лектор, Алматинский технологический университет, Алматы, Казахстан, e-mail: gani1212@bk.ru;

Компан С.И. - Алматинский технологический университет, Алматы, Казахстан, e-mail: a.sv21@mail.ru;

Казангапова Б.А. - ассоциированный профессор, Алматинский технологический университет, Алматы, Казахстан, e-mail: kbayana@mail.ru.

Information about the authors

Bukenova I. - Master of Technical Sciences, lecturer, Almaty Technological University, Almaty, Kazakhstan, e-mail: ibukenowa@mail.ru;

Bukenov G. - Master of Technical Sciences, lecturer, Almaty Technological University, Almaty, Kazakhstan, e-mail: gani1212@bk.ru;

Kompan S. - Almaty Technological University, Almaty, Kazakhstan, e-mail: a.sv21@mail.ru;

Kazangapova B. - Associate Professor, Almaty Technological University, Almaty, Kazakhstan, e-mail: kbayana@mail.ru.