УДК 004.89

ИНТЕГРАЦИЯ СИСТЕМ ГЕНЕРАЦИИ НА ОСНОВЕ ПОИСКА И СЕМАНТИЧЕСКОГО КЭШИРОВАНИЯ ДЛЯ ПОВЫШЕНИЯ ДОСТОВЕРНОСТИ ПРЕДМЕТНО-ОРИЕНТИРОВАННЫХ LLM

А.В. Шкрабов
Студент кафедры электронной техники и технологии БГУИР
sashashkrabov04@gmail.com

Н.А. Резников
Студент кафедры электронной техники и технологии БГУИР
mine.turtle467@gmail.com

С. К. Дик
Заведующий кафедрой инженерной и компьютерной графики БГУИР,
канд. физ.-мат. наук, доцент
sdick@bsuir.by

В.М. Бондарик
Декан факультета доуниверситетской подготовки и профессиональной ориентации БГУИР, канд. техн. наук, доцент
bondarik@bsuir.by

И.И. Ревинская
Старший преподаватель кафедры электронной техники и технологии БГУИР
inna_revinskaya@bsuir.by

А.В. Шкрабов

Студент кафедры электронной техники и технологии Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем.

Н.А. Резников

Студент кафедры электронной техники и технологии Белорусского государственного университета информатики и радиоэлектроники. Область научного интереса – компьютерная инженерия, программирование.

С.К. Дик

Заведующий кафедрой инженерной и компьютерной графики Белорусский государственный университет информатикии радиоэлектроники, кандидат физико-математических наук, доцент. Окончил Минский радиотехнический институт по специальности «Радиотехника», руководит научными исследованиями в области лазерной медицины.

В.М Бондарик

Окончил Минский радиотехнический институт по специальности «Конструирование и производство радиоаппаратуры». Кандидат технических наук, доцент кафедры электронной техники и технологии. Область интересов медицинская электронная техника, проектирование ультразвуковых медицинских электронных систем и внедрение дистанционных образовательных технологий.

И.И. Ревинская

Окончила Белорусский государственный университет информатики и радиоэлектроники. Старший преподаватель кафедры электронной техники и технологии. Область научного интереса – медицинская электроника и обработка медицинских сигналов.

Аннотация. В статье приведено решение проблемы сокращения галлюцинаций больших языковых моделей (*LLM*) при работе с высоконагруженными распределенными базами данных в архитектуре *RAG*. Анализируется влияние системных факторов (сетевых задержек, асинхронной консистентности, семантического шума) на фактологическую точность генерации. В качестве решения рассмотрены инновационные подходы: децентрализованная архитектура *DRAG*, многоисточниковая кросс-верификация *MEGA-RAG* на базе графов знаний, интеллектуальное семантическое кэширование и стратегии многофазного ранжирования. Установлено, что комплексное внедрение данных методов сводит к минимуму фактологическую манипуляцию, эффективно отсеивает «жесткие негативы» и оптимизирует баланс между барьером и скоростью вывода без ресурсоемкого дообучения модели.

Ключевые слова: большие языковые модели (*LLM*), генерация, дополненный поиск (*RAG*), галлюцинации нейросетей, распределенные базы данных, *DRAG*, *MEGA-RAG*, семантическое кэширование, графы знаний.

Введение. Фундаментальный принцип работы современных больших языковых моделей (LLM) основывается на механизме авторегрессионного предсказания следующего токена, при котором целевая функция нейронной сети оптимизируется на минимизацию потерь при воспроизведении вероятностных и дистрибутивных статистик, полученных из больших массивов обучающих данных. Однако эта же архитектурная особенность порождает уязвимость: процесс генерации происходит без поддержания явного, детерминированного соответствия объективной внешней реальности или формальным базам фактов. Модель оперирует исключительно в латентном пространстве выученных вероятностей, что неизбежно приводит к феномену конфабуляции, или галлюцинаций – генерации правдоподобной, но фактически недостоверной или логически некорректной информации. Внедрение таких систем в важные домены, включая клиническую медицину, юриспруденцию, финансовый аудит и управление промышленными системами, наталкивается на жесткие регуляторные барьеры, которые требуют высоких уровней прозрачности, интерпретируемости и абсолютной фактологической точности.

Для преодоления ограничений, связанных со статичностью знаний, "замороженных" в весовых коэффициентах параметров модели на этапе предварительного обучения, исследовательским сообществом была разработана концепция генерации, дополненной

поиском (RAG). В рамках этого подхода на первом этапе пользовательский запрос векторизуется и используется для извлечения релевантных документов, параграфов или структурированных фрагментов данных из внешней базы знаний. На втором этапе эти извлеченные фрагменты интегрируются в исходный контекст (промт) вместе со специализированными системными инструкциями, направляя большую языковую модель на использование именно этой актуальной информации при формулировании итогового ответа. Таким образом, RAG позволяет расширить фактологическую базу модели до неограниченных масштабов, предоставляя доступ к специализированным закрытым корпоративным доменам знаний без необходимости инициации экспоненциально дорогих процессов дообучения всей нейронной сети.

Однако по мере масштабирования подобных систем до корпоративного уровня архитектурный фокус неизбежно смещается в сторону применения высоконагруженных распределенных баз данных для хранения миллиардов векторных представлений (эмбеддингов) и ассоциированных текстовых фрагментов. Распределенные среды характеризуются физическим разделением узлов хранения, необходимостью сложной сетевой маршрутизации и применением компромиссных моделей консистентности данных. В таких условиях к классическим причинам возникновения галлюцинаций самой LLM добавляются системные сбои конвейера извлечения: сетевые задержки могут приводить к рассинхронизации компонентов, а проблемы согласованности – к извлечению устаревших, поврежденных или противоречивых фрагментов знаний. Когда генеративная модель получает на вход гетерогенный или внутренне противоречивый контекст из распределенной базы, ее механизмы внимания пытаются синтезировать единый ответ, что с высокой долей вероятности приводит к комбинированным галлюцинациям. Следовательно, проблема снижения галлюцинаций в масштабируемых RAG-системах перестает быть исключительно лингвистической или нейросетевой задачей; она трансформируется в комплексную проблему инженерии распределенных систем, требующую глубокого анализа взаимодействия между алгоритмами векторного поиска, топологией сети и механизмами внимания больших языковых моделей.. [1].

Многомерная таксономия галлюцинаций и их геометрические сигнатуры в латентных пространствах. Чтобы эффективно бороться с галлюцинациями, необходимо детально разобрать само это понятие галюцинаций, поскольку сегодня этим общим термином называют совершенно разные системные сбои. Ситуация, при которой языковая модель полностью игнорирует предоставленные в контекстном окне релевантные документы из распределенной базы, кардинально отличается от ситуации, когда модель генерирует абсолютно несуществующие сущности на основе шума, или от случая, когда она предоставляет искаженные, но правдоподобные детали о реально существующих концепциях. Данные сбои имеют различные корневые причины, разные последствия для конечного пользователя и требуют применения принципиально разных стратегий алгоритмической коррекции. Комплексный анализ причинно-следственных связей требует рассмотрения всего жизненного цикла разработки большой языковой модели, начиная от этапа сбора данных и архитектурного проектирования, заканчивая этапами тонкой настройки и инференса в *RAG*-конвейере.

В контексте анализа автономных ИИ-агентов и систем, постоянно взаимодействующих с распределенными базами данных, современная можно выделить несколько специфических категорий галлюцинаций:

1) галлюцинации рассуждения напрямую влияют на способность модели анализировать поведение и принимать последовательные, логически обоснованные решения при многошаговом поиске в графах или реляционных таблицах;

2) галлюцинации выполнения, когда модель некорректно применяет инструменты или формирует синтаксически неверные запросы (например, ошибочные *SQL*-запросы к распределенным базам);

3) галлюцинации восприятия возникают при некорректной интерпретации контекста или мультимодальных данных;

4) галлюцинации запоминания связаны с конфликтом между знаниями, инкапсулированными в весах модели, и знаниями, извлеченными из векторной базы;

5) галлюцинации коммуникации проявляются в виде непоследовательного диалогового взаимодействия с пользователем или другими агентами в сети.

В ходе анализа автономных ИИ-агентов и систем, постоянно взаимодействующих с распределенными базами данных, выделены основные причины возникновения галлюцинаций (таблица 1) [2].

Таблица 1. Основные причины и механизмы возникновения галлюцинаций в *LLM*

Причина	Фактор уязвимости	Механизм возникновения галлюцинации и влияние на генерацию
Качество обучающих данных	Зашумленность и неполнота	Внедрение ошибочных паттернов в параметрическую память модели, что приводит к некорректным ответам даже при наличии правильного контекста.
Смещение данных	Дефицит репрезентативного разнообразия	Провоцирует имитационные конфабуляции, когда модель чрезмерно экстраполирует узкие доменные корреляции на универсальные запросы.
Временная деградация	Использование устаревших (статичных) корпусов	Генерация фактологических ошибок при ответах на запросы о динамически меняющемся мире, требующая жесткого перекрытия через механизмы *RAG*.
Архитектура и инференс	Природа авторегрессионного декодирования	Оптимизация вероятностных распределений без привязки к семантической истине (отсутствие детерминированного обоснования).

Различные классы галлюцинаций больших языковых моделей поддаются математической дифференциации на основе анализа их топологического поведения на многомерной единичной гиперсфере векторных вложений, обозначаемой как C^{d-1}. В качестве математического инструмента для такого анализа был предложен Индекс семантического обоснования, который количественно измеряет степень вовлеченности сгенерированного ответа во взаимодействие с предоставленным контекстом. Механика данного индекса базируется на вычислении сложных отношений угловых расстояний между векторами токенов ответа и векторами исходных документов на единичной гиперсфере [3]. Индекс семантического обоснования достигает высочайшего уровня дискриминации (величина эффекта d Коэна находится в интервале 0,92–1,28) при обнаружении феномена недобросовестности – специфического типа галлюцинации, при котором модель генерирует текст, семантически ортогональный извлеченному контексту. Математический метод (Индекс семантического обоснования) разделяет «нормальный ответ» и «галлюцинацию» настолько четко, что они выглядят как абсолютно разные категории объектов. Система не просто «слегка замечает» ошибку, она видит гигантский разрыв между достоверным фактом и выдумкой нейросети, который позволяет изолировать такие галлюцинации со 100% уверенностью. Понимание этих геометрических характеристик открывает путь к созданию превентивных фильтров на уровне логического вывода: если система векторного поиска фиксирует, что формируемый ответ начинает отклоняться от гиперсферического кластера извлеченных документов, генерация может быть прервана или скорректирована до того, как галлюцинация будет транслирована конечному пользователю [4].

Влияние сетевых задержек и асинхронной консистентности на фактологическую точность. В процессе эволюции векторного поиска от изолированных прототипов к промышленным распределенным архитектурам проявляется существенный недостаток *RAG*-систем, который заключается в их низкой устойчивости к системным сбоям и аномалиям в работе баз данных. Качество генерируемого текста и вероятность возникновения галлюцинаций зависит от чистоты, релевантности и своевременности документов, предоставляемых модулем извлечения. В архитектуре распределенных баз данных одним из наиболее разрушительных факторов является сетевая задержка – временной интервал, необходимый для передачи пакетов данных между физически разнесенными вычислительными узлами кластера. В масштабной распределенной среде, особенно построенной на принципах микросервисов или глобального гео-шардирования, данные хранятся фрагментированно на множестве серверов. Любая комплексная операция, требующая доступа к различным сегментам знаний или обновления векторных индексов, неизбежно сопряжена с многократным межсетевым взаимодействием.

В распределенных архитектурах *RAG* возрастание сетевых задержек выступает фактором, лимитирующим производительность системы. Данное воздействие проявляется в двух аспектах:

1) извлечение семантически релевантных чанков из гетерогенных географических зон приводит к накоплению критической задержки (кумулятивной латентности), что исключает возможность использования синхронных методов поиска в контуре генерации ответа;

2) высокие сетевые задержки препятствуют реализации транзакций с гарантиями строгой консистентности.

Процессы обновления базы знаний (векторного индекса) требуют координации между узлами распределенной системы; превышение пороговых значений латентности делает синхронный консенсус недостижимым, вынуждая применять парадигму согласованности в конечном счете.

Режим согласованности в конечном счете в распределенных БД характеризуется наличием временного окна, в течение которого узлы содержат противоречивые данные. Это приводит к состоянию гонки при конкурентном обновлении записей, следствием чего является деградация векторных индексов. Для *RAG*-пайплайна данная деградация оборачивается извлечением семантически несогласованных фрагментов, содержащих конфликтующие факты. Поскольку архитектура *LLM* не предполагает встроенного логического фактчекинга, механизмы внимания нивелируют противоречия путем генерации синтетических, статистически правдоподобных, но ложных конструкций, что идентифицируется как фактологическая галлюцинация. Нивилирование данной проблемы требует применения компенсаторных архитектурных паттернов: многоуровневого кэширования на границе сети, репликации с учетом локальности запросов и динамической оптимизации маршрутизации.

Децентрализованная парадигма. Классические централизованные *RAG*-системы, демонстрируя высокую эффективность в лабораторных условиях, при развертывании в реальных высоконагруженных доменах сталкиваются с ограничениями масштабируемости и обеспечения конфиденциальности данных. Особую остроту данная проблема приобретает в чувствительных областях, таких как прецизионного здравоохранения, требует агрегации высокочувствительных персональных биометрических и медицинских данных пациентов. Централизованные монолитные базы знаний вступают в противоречие с регуляторными требованиями суверенитета данных, а поддержание их актуальности в динамически изменяющихся условиях сопряжено с экономически неоправданно высокими затратами.

Разрешением указанного архитектурного тупика выступает парадигма распределенной генерации с дополненным поиском (*DRAG*), реализующая переход от облачных монолитов к одноранговым сетям периферийных вычислений. В архитектуре DRAG каждый участник сети выступает автономным узлом, обладающим локальной базой

знаний, приватным экземпляром языковой модели и модулем защищенной коммуникации. Обработка информационного запроса структурирована как пятиэтапный конвейер.

1. Инициирующий узел силами локальной малой языковой модели (*sLLM*) выполняет парсинг запроса и извлекает имплицитные тематические сигнатуры. На основе вычисления косинусного сходства между эмбеддингом запроса и локальным графом знаний принимается решение о достаточности локального контекста. При низких значениях сходства или негативных эвристиках удовлетворенности пользователя инициируется переход к стадии внешнего поиска.

2. Для навигации в децентрализованной сети применяется алгоритм тематически-ориентированного случайного блуждания. В отличие от лавинной маршрутизации, *TARW* использует аналитический потенциал локальной *LLM* для направленного поиска: модель идентифицирует узлы-кандидаты на основе кэшированных таблиц экспертизы, направляя запрос только тем пирам, чья историческая релевантность в данной предметной области подтверждена.

3. Узлы-доноры не экспортируют сырые текстовые данные. Вместо этого локальная *LLM* генерирует «отфильтрованные пользователем фрагменты знаний», применяя маскировку конфиденциальных паттернов. Данный механизм гарантирует абсолютный суверенитет исходных данных, транслируя их в семантически обогащенный, но деидентифицированный формат.

4. Инициирующий узел агрегирует полученные деидентифицированные фрагменты с собственным локальным контекстом, формируя расширенное семантическое пространство для финальной генерации ответа.

5. Сгенерированный ответ кэшируется, а метрики репутации пиров-доноров обновляются. Это обеспечивает адаптивную маршрутизацию будущих запросов, позволяя обходить деградировавшие или высоконагруженные узлы, что повышает общую отказоустойчивость распределенной системы.

Экспериментальная оценка алгоритма *TARW* демонстрирует его эффективность в распределенной архитектуре *DRAG*: на бенчмарке *MMLU* достигнута точность *Exact Match* 83,90% (против 85,73% у централизованного *RAG*) при сокращении межсетевого трафика до 6,87 сообщений на запрос (против 10,91 при *flooding*). Главная проблема распределенного подхода – гетерогенность узлов и риск контаминации данных – компенсируется внедрением криптографической репутационной системы и взвешенного ранжирования пиров по исторической валидности [6].

Многоисточниковая кросс-верификация и графовые структуры в архитектуре *MEGA-RAG*. Минимизация фактологических обобщений в самых важных предметных областях (таких как клиническая медицина, фармакология и юриспруденция) обуславливает необходимость разработки архитектуры с нулевой толерантностью к конфабуляциям. Подобный уровень надежности принципиально недостижим в рамках базовых парадигм *RAG*, вне зависимости от степени их децентрализации. Обычные генеративные агенты часто не способны распознать логические противоречия при агрегации нескольких фрагментов текста, что делает их непригодными для клинического использования. Для решения данных проблем был предложен специализированный фреймворк *MEGA-RAG* (Поисковая дополненная генерация с уточнением ответов на основе нескольких доказательств), представляющий собой комплексную четырехэтапную архитектуру, направленную на минимизацию галлюцинаций.

В отличие от стандартной парадигмы, распределяющейся исключительно на плотных векторных индексах, *MSER* реализует ортогональную последовательную агрегацию гетерогенных данных из трех структурно различных распределенных хранилищ. К ним относятся полнотекстовые базы рецензируемой биомедицинской литературы *PubMed*, обрабатываемые с помощью плотных векторных библиотек *FAISS*; нормативная база

данных *IRIS* Всемирной системы здравоохранения, поиск в которой осуществляется с применением лексического алгоритма организации *BM25* [7].

Интеграция графических баз данных (*Graph-RAG*) является одним из наиболее эффективных методов подавления галлюцинаций в больших языковых моделях (*LLM*). Векторные системы строят связи на основе вероятностного математического подобия, тогда как графы знаний оперируют детерминированными ребрами, отражающими точные онтологические отношения между сущностями (например, в логической триаде «симптом – препарат – побочный эффект»). Подобная топология позволяет делать логические выводы непосредственно на уровне системы хранения данных, отслеживая эту задачу генеративной модели, динамической стохастической конфабуляции. Отсутствие иского факта в графе приводит к возврату пустого распространения, которое инициирует правильный отказ модели от формирования ответа и блокирует генерацию артефактов на основе дополнительных векторных соотношений. Последующая обработка в рамках фреймворка включает в себя три стадии глубокого анализа. Второй компонент – модуль генерации вариативных ответов (*DPAG*) – формирует массив кандидатных решений методом сэмплирования, после чего к ним применяются вычислительно ресурсоемкие модели кросс-энкодеров. Совместная обработка запроса и кандидатного документа через пошаговый трансформер обеспечивает выявление тонких семантических связей, исключающих задержки биэнкодерам. На третьем этапе модуль семантико-эвиденциального спорта (*SEAE*) использует метрику *BERTScore* для математической квантификации степени обоснованности каждого сгенерированного токена полученной доказательной базой. Основным инновационным механизмом выступает четвертый модуль принципа расхождений и саморазъяснения (*DISC*). При падении метрики SEAE ниже установленного порогового значения данный компонент не прерывает генерацию, анализ проводит семантические дивергенции между массивом потенциальных ответов, синтезирует вторичный корректирующий запрос к хранилищам данных и реализует процесс направленного редактирования формируемых знаний. Экспериментальные замеры в задачах медицинского профиля убедительно доказывают превосходство *MEGA-RAG* над изолированными большими языковыми моделями, дообученными системами вроде *PubMedGPT* и базовыми реализациями *RAG*. Архитектура достигает высокой точности и прозрачности, гарантируя, что любые фактологические утверждения подкреплены перекрестно проверенными источниками (таблица 2) [8].

Таблица 2. Сравнительная эффективность моделей

Тип модели	Точность	Снижение уровня галлюцинаций
PubMedBERT (изолированная модель)	Низкая (частые фабрикации)	Базовая линия (0%)
Standalone LLM (GPT-4)	Средняя	Около 15%
Стандартный векторный *RAG*	Умеренная	Около 25%
MEGA-RAG	Высокая	Свыше 40% относительно стандартов

Интеллектуальное семантическое кэширование: компромисс между задержкой и ложными срабатываниями. Развертывание масштабируемых систем *RAG* в производственной среде сопряжено с постоянной борьбой против двух главных врагов архитектуры: стоимости обращения к большим языковым моделям и задержек логического вывода, разрушающих пользовательский опыт. Для приложений, характеризующихся высокой плотностью повторяющихся запросов с небольшими вариациями формулировок (например, чат-боты клиентской поддержки, интеллектуальные системы анализа технической документации), классические методы префиксного кэширования оказываются недостаточными. Механизмы кэширования промптов, реализованные на уровне провайдеров (таких как *Microsoft Azure* или *OpenAI*), функционируют путем сохранения

тензоров «ключ-значение» механизмов внимания для начальных токенов запроса. Этот подход требует идеального, побайтового совпадения префиксной части запроса. Аналогичным образом, традиционное кэширование типа ключ-значение (базирующееся на хешировании строк) не способно обрабатывать естественный язык: запросы «Какова ваша политика возврата?», «Как я могу вернуть товар?» и «Могу ли я отправить это обратно?» будут восприниматься системой как абсолютно разные ключи, инициируя три отдельных, дорогостоящих и медленных прохода через всю цепь извлечения и генерации.

Концептуальным решением этой проблемы является внедрение интеллектуального слоя семантического кэширования. Семантическое кэширование оперирует не синтаксисом, а исключительно математическим смыслом и интенцией запроса. Алгоритм функционирует следующим образом: поступающий пользовательский запрос конвертируется легковесной моделью в векторное представление. Затем этот вектор используется для поиска в высокоскоростной резидентной базе данных, где хранятся эмбеддинги исторических запросов, сопряженные с ранее сгенерированными эталонными ответами. Поиск осуществляется через вычисление косинусного сходства между векторами в многомерном пространстве. Если показатель сходства превышает жестко заданный порог уверенности, система мгновенно возвращает кэшированный ответ, полностью обходя этапы векторного поиска по базе документов и генерации ответа тяжелой *LLM*. Внедрение этого паттерна обеспечивает феноменальный прирост производительности: время отклика может сократиться с ~6.5 секунд (полный цикл *RAG*) до 100 миллисекунд (извлечение из кэша) – 65-кратное ускорение, сопровождающееся пропорциональным падением операционных расходов.

Однако внедрение семантического кэша таит в себе скрытую угрозу, известную в индустрии как «кризис ложных срабатываний». Векторное пространство не идеально; при некорректной настройке гиперпараметров система начинает выдавать уверенные, семантически близкие, но фактически неверные ответы из кэша, заменяя фактологическую точность на скорость. Для нейтрализации этого существенного риска необходимо перейти от простого тюнинга порогов к многослойному архитектурному дизайну. Фундаментальным принципом предотвращения ложных срабатываний является «принцип наилучшего кандидата». Данный принцип постулирует, что вместо инкрементального заполнения кэша «на лету», производственная система должна быть предварительно инициализирована «Золотым стандартом» – каноническими ответами на вопросы, покрывающими основные предметные домены. Существенно важно, что вместе с золотым стандартом в базу внедряются «стратегические дистракторы» – синтетические запросы, которые семантически похожи на канонические, но требуют принципиально иного ответа. Идеальное эмпирическое соотношение составляет 3 дистрактора на 1 эталонный запрос. Наличие таких контрольных точек в латентном пространстве позволяет алгоритму предельно точно вычислять границы смысловых кластеров, предотвращая ошибочное объединение разных концепций. Оптимизация порога срабатывания является не менее значимым фактором и сильно зависит от выбора базовой модели эмбеддингов, так как плотность распределения векторов у разных архитектур кардинально отличается.

Дальнейшее снижение уровня ложных срабатываний (ниже 5%) требует реализации многовекторной архитектуры, при которой создаются изолированные индексные пространства для различных аспектов одного запроса. Кроме того, внедрение легковесных моделей кросс-энкодеров на этапе валидации кандидата из кэша позволяет провести глубокий анализ взаимосвязей между входным запросом и кэшированным эталоном до отправки ответа пользователю, фактически исключая семантические галлюцинации на уровне кэша [9].

Многофазное ранжирование и фильтрация семантического шума. Управление распределенными цепями *RAG* требует жесткой балансировки между задержкой вывода и выходом (компромисс задержки и точности). Линейное расширение контекстного окна или

увеличение выбора (параметр *Top-K*) перегружает механизмы внимания и провоцирует галлюцинации типа «потери в середине». Проблема решается внедрением методик многофазного ранжирования, декомпозирующего процесса определения. Первичный обзор массива данных осуществляется высокопроизводительными лексическими (*BM25*) и векторными (*ANN*) алгоритмами, для чего применяется гибридное переранжирование лучших кандидатов. На конечном этапе ресурсоемкие кросс-энкодеры проводят прецизионный семантический анализ ограниченной выборки, в результате чего определяется релевантность без критической деградации латентной системы. Существенным источником происхождения конфабуляций является наличие в выборке «жестких негативов» – документов с высоким поверхностным сходством, но фактологически неверным добавлением. Нейтрализация данного семантического шума требует применения комплексных алгоритмов на базе методов атрибуции. В условиях ограниченного доступа к радиусам моделей в распределенных базах рекомендуется использовать атрибуции «черного ящика» с привлечением альтернативных фильтрующих моделей, заменяющих градиентный анализ («белый ящик»). Подобная комплексная оптимизация шумоподавления и ранжирование серьезности задач трансформации естественного языка в *SQL*-запросы (*Text-to-SQL*). В отличие от базовых архитектур с использованием схемы фиксированного количества (*fixed-k*), внедрение алгоритмов динамической иерархической кластеризации позволяет адаптивно определять релевантный объем таблиц, отсекающий шум на этапе формирования контекста. Синергия данной ситуации с итеративным обогащением предложения реальных записей повышения метрику *F*1 извлечения схем с 0,79 до 0,88 и повышает точность выполнения запросов на 11%, полностью выполняя необходимость ресурсоемкого дообучения (тонкой настройки) модели [10].

Заключение. Интеграция больших языковых моделей с распределенными корпоративными базами данных в рамках структуры *RAG* актуализирует фундаментальную проблему фактологической достоверности искусственного интеллекта. Сетевые задержки, компромиссные модели асинхронной консистентности и присутствие семантического шума трансформируют функцию подавления галлюцинаций из-за лингвистической в комплексной инженерной проблеме. Преодоление выявленных ограничений требует концептуального отказа от базовых парадигм линейного векторного поиска с использованием глубоко эшелонированных, многофазных архитектурных обработок данных. Анализ современных подходов доказывает эффективность децентрализованных систем класса *DRAG*, которые за счет алгоритмов направленного топологического блуждания решают проблемы масштабируемости сети и суверенитета конфиденциальных данных. В важнейших областях, требующих нулевой толерантности к конфабуляциям, архитектурными решениями является многофакторная кросс-верификация (фреймворк *MEGA-RAG*). Интеграция вероятностного векторного определения со строгой онтологической детерминированностью графов знаний (*Graph-RAG*) позволяет снизить уровень фактологических сбоев. Вычислительная рентабельность таких тяжеловесных процессов внедрения уровней интеллектуального семантического кэширования с использованием ложных программ и методик иерархического многофазного ранжирования, которые минимизируют латентность вывода и эффективно фильтруют «жесткие ошибки».Обеспечение надежности генеративных систем в высоконагруженных распределенных средах не учитывает экстенсивного масштабирования параметров языковых моделей и их ресурсоемкого обучения, а следовательно, глубокой алгоритмической оптимизации конвейера извлечения и верификации контекста. Последующая эволюция инженерных знаний в данной области будет базироваться на развитии гибридных методов шумоподавления, декомпозиции простых запросов для итеративного фактчекинга и внедрения математически обоснованных метрик геометрического контроля семантики на всех этапах логического вывода.

Список литературы

[1] Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks //Advances in neural information processing systems. – 2020. – Т. 33. – С. 9459-9474. – URL: https://arxiv.org/pdf/2005.11401 (дата обращения 27.02.2026).

[2] Sajjadi Mohammadabadi, S. M.; Kara, B. C.; Eyupoglu, C.; Karakus, O. A Survey on Hallucination in Large Language Models: Definitions, Detection, and Mitigation. Preprints 2025, 2025100540. – URL: https://doi.org/10.20944/preprints202510.0540.v1 (дата обращения 28.02.2026)

[3] Marín J. A Geometric Taxonomy of Hallucinations in LLMs //arXiv preprint arXiv:2602.13224. – 2026. – URL: https://arxiv.org/html/2602.13224v1 (дата обращения 27.02.2026).

[4] Marín J. Semantic grounding index: Geometric bounds on context engagement in RAG systems //arXiv preprint arXiv:2512.13771. – 2025. – URL: https://arxiv.org/pdf/2512.13771 (дата обращения 27.02.2026).

[5] Gao Y. et al. Retrieval-augmented generation for large language models: A survey //arXiv preprint arXiv:2312.10997. – 2023. – Т. 2. – №. 1. – С. 32. – URL: https://www.semanticscholar.org/reader/46f9f7b8f88f72e12cbdb21e3311f995eb6e65c5 (дата обращения 27.02.2026).

[6] Xu C. et al. Distributed retrieval-augmented generation //arXiv preprint arXiv:2505.00443. – 2025. – URL: https://arxiv.org/html/2505.00443v1 (дата обращения 27.02.2026).

[7] Xu S. et al. MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health //Frontiers in Public Health. – 2025. – Т. 13. – С. 1635381. – URL: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2025.1635381/full (дата обращения 27.02.2026).

[8] Edge D. et al. From local to global: A graph rag approach to query-focused summarization //arXiv preprint arXiv:2404.16130. – 2024. – URL: https://arxiv.org/pdf/2404.16130 (дата обращения 27.02.2026).

[9] Bang F. Gptcache: An open-source semantic cache for llm applications enabling faster answers and cost savings //Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023). – 2023. – С. 212-218. – URL: https://openreview.net/pdf?id=ivwM8NwM4Z (дата обращения 27.02.2026).

[10] Bozdemir M., Bilgin M. Schema Retrieval with Embeddings and Vector Stores Using Retrieval-Augmented Generation and LLM-Based SQL Query Generation //Applied Sciences. – 2026. – Т. 16. – №. 2. – С. 586. – URL: https://www.mdpi.com/2076-3417/16/2/586 (дата обращения 27.02.2026).

Авторский вклад

Авторы внесли равный вклад в написание статьи

INTEGRATION OF SEARCH-BASED GENERATION SYSTEMS AND SEMANTIC CACHING TO INCREASE THE RELIABILITY OF DOMAIN-ORIENTED LLMS

N.A. Reznikov
student of the Department of Electronic Technique and Technology of BSUIR

A.V. Shkrabov
student of the Department of Electronic Technique and Technology of BSUIR

S. K. Dzik
Chair of the Department of Engineering and Computer Graphics, PhD, Associate Professor

V.M. Bondarik
Dean of the Faculty of Pre-University Training and Career Guidance
Associate Professor of the Department of Electronic Engineering and Technology

I.I. Revinskaya
Senior Lecturer at the Department of Electronic Technique and Technology

Abstract. The paper discusses an approach to increasing the robustness of neural network models used in eye footage segmentation to typical image distortions using targeted data augmentation. Using the example of the winner of the OpenEDS Semantic Segmentation Challenge 2019 – the RitNet model (based on U-Net and DenseNet), it is shown how the addition of synthetic artifacts (structured starbursts, Gaussian blur, random lines and shifts), characteristic of images in VR / AR glasses, made it possible to achieve high accuracy with an extremely small model size. This approach emphasizes the importance of adapting training data to real-world application conditions to improve the reliability of models in computer vision tasks.

Keywords: big data, convolutional neural networks, segmentation, model robustness, infrared cameras, data augmentation, eye detection systems