АРХИТЕКТУРА RETRIEVAL-AUGMENTED GENERATION КАК ОСНОВА ПОСТРОЕНИЯ ЛОКАЛЬНЫХ ИИ-СИСТЕМ С КОРПОРАТИВНЫМИ БАЗАМИ ЗНАНИЙ

Арефин В. А.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: arefin.vlad@gmail.com

Целью работы является исследование архитектуры Retrieval-Augmented Generation (RAG) как основы построения локальных интеллектуальных систем, взаимодействующих с корпоративными базами знаний. Рассмотрены ключевые компоненты архитектуры и принципы интеграции механизма извлечения знаний с языковыми моделями. Проведён анализ преимуществ RAG в контексте корпоративных информационных систем, включая повышение точности и достоверности ответов, снижение вероятности генерации ложных данных и обеспечение актуальности знаний без переобучения модели. Определены основные направления развития и практические аспекты внедрения архитектуры RAG в локальных ИИ-решениях.

I. Введение

Современные большие языковые модели (Large Language Models, LLM) демонстрируют значительный прогресс в генерации связного текста и ведении диалога, однако они имеют фундаментальные ограничения при работе с внутренними корпоративными данными. Большинство моделей обучены на открытых наборах данных и не обладают механизмом динамического обновления знаний. Это делает их использование в корпоративных сценариях неэффективным без дополнительной архитектурной поддержки.

Метод Retrieval-Augmented Generation (RAG) решает данную проблему, совмещая генерацию и поиск релевантных данных [1]. Перед формированием ответа система выполняет извлечение необходимых фрагментов из базы знаний, что повышает точность и обоснованность результатов, снижает эффект «галлюцинаций» и адаптирует модель под специфику организации [2, 5]. Подход RAG активно применяется при построении корпоративных интеллектуальных систем, чат-ботов, экспертных ассистентов и платформ поддержки принятия решений.

II. Архитектура RAG

Архитектура RAG обычно включает три ключевых этапа: (1) подготовку и индексацию базы знаний (chunking, векторизация); (2) извлечение релевантных документов по запросу (retrieval); (3) генерацию ответа с учётом найденных фрагментов (generation) [6]. На практике также используются дополнительные механизмы: reranking, фильтрация по достоверности, комбинирование источников и последующая верификация.

Подготовка корпоративной базы знаний

Корпоративная база знаний может включать регламенты, технические отчёты, инструкции, записи совещаний, внутренние чаты и инцидентные отчёты. Для эффективного применения RAG необходимо провести очистку, структурирование и семантическую разметку данных. Формирование векторного индекса на основе эмбеддингов обеспечивает быстрый поиск релевантных фрагментов, а сохранение метаданных позволяет контролировать источники и версии документов. Особое внимание уделяется вопросам безопасности, разграничению доступа и контролю за актуальностью данных.

Этап извлечения и генерации

На этапе retrieval пользовательский запрос преобразуется в векторное представление и сопоставляется с индексом знаний. Далее выбираются K наиболее релевантных фрагментов, которые подаются вместе с запросом в генеративную модель. Эта модель формирует ответ, опираясь не только на параметры своего обучения, но и на актуальные корпоративные данные [7]. Такой подход обеспечивает синтез свежих и достоверных ответов даже при ограниченном обучении основной LLM.

Архитектурный паттерн и инфраструктура

Типовая инфраструктура RAG включает:

- систему хранения и индексации документов (векторные, графовые или гибридные БД);
- модуль поиска и ранжирования релевантных фрагментов;
- генеративную модель с продуманным prompt-инжинирингом;
- подсистему контроля качества и интерпретации ответов (fact-checking, логирование источников);
- интерфейс пользователя — чат-бот, API или аналитический портал.

Для корпоративных сценариев важна возможность локального развёртывания (on-premises) с соблюдением требований безопасности

и приватности. Согласно аналитическим данным, более половины компаний, внедряющих решения на базе ИИ, рассматривают архитектуру RAG как основную при работе с внутренними данными [1].

III. Применение в локальных корпоративных системах

Кейс: корпоративный чат-бот

В крупных организациях объём внутренних документов может достигать десятков тысяч файлов, включая нормативные акты, регламенты, методические указания, отчётные формы и переписку сотрудников. Такие массивы данных являются ценным источником корпоративных знаний, однако их использование традиционными поисковыми средствами ограничено.

Применение архитектуры RAG позволяет создать интеллектуального чат-бота, который при каждом пользовательском запросе выполняет поиск релевантных документов, анализирует их содержание и формирует ответ с указанием конкретных источников [3, 6]. В результате система становится полноценным интеллектуальным помощником, способным интерпретировать контекст и связывать разрозненные сведения.

Подобный подход значительно повышает доверие пользователей, сокращает время на поиск информации и обеспечивает соответствие ответов актуальным нормативам и внутренним политикам организации.

Преимущества и вызовы

Применение RAG в корпоративной среде даёт ряд стратегических преимуществ:

- снижение вероятности генерации недостоверных или вымышленных ответов благодаря обращению к проверенным источникам;
- актуализация знаний без повторного обучения модели, что снижает затраты на сопровождение;
- возможность внедрения в изолированной корпоративной инфраструктуре с соблюдением политик безопасности и конфиденциальности;
- повышение прозрачности работы системы и доверия со стороны пользователей за счёт цитирования источников и ведения журналов обращений;

Успешное внедрение RAG требует решения ряда технологических и организационных задач. Ключевыми вызовами являются обеспечение качества индексации и полноты охвата базы знаний, масштабируемость архитектуры при росте объёмов данных, а также поддержка многоязычных документов и сложных форматов (PDF, таблицы, графики).

IV. Выводы

Архитектура RAG формирует основу для построения надёжных и адаптивных корпоративных ИИ-систем [4]. Её применение позволяет соединить преимущества больших языковых моделей и актуальных баз знаний, обеспечивая обоснованность и свежесть ответов. Перспективными направлениями развития являются интеграция мультимодальных данных, исследование самонастраивающихся retrieval-модулей, а также оптимизация производительности при обработке больших объёмов корпоративной информации [5, 2].

Список литературы

1. Yee, L., Chui, M., Roberts, R., Pometti, M., Wollner, P. What is Retrieval-Augmented Generation (RAG)? // McKinsey & Company Explainer. – 2024.
2. Zhang, H., Chen, M., Li, T. Optimizing Retrieval-Augmented Generation Pipelines for Domain-Specific Knowledge Bases // Proc. IEEE Int. Conf. on Artificial Intelligence and Knowledge Engineering (AIKE). – 2023. – P. 215-222.
3. Нестеренков, С. Н., Красовский, В. Ю., Баяк, Е. И. Искусственный интеллект и персональные ассистенты в управлении проектами // Труды БГУИР. – 2024. – № 2 (140). – С. 35-42.
4. Нестеренков, С. Н., Байчик, С. А., Голубович, Ю. И. Искусственный интеллект в процессах обучения и развития персонала // Информационные технологии и системы: материалы конф. – Минск: БГУИР, 2024. – С. 112-115.
5. Gupta, S., Ranjan, R., Singh, S. N. A Comprehensive Survey of Retrieval-Augmented Generation: Evolution, Current Landscape and Future Directions // arXiv preprint arXiv:2410.12837. – 2024.
6. Martinez, A., Bae, J., Qian, Y. Towards Reliable Corporate AI Assistants: Improving Contextual Reasoning in RAG Systems // Proceedings of the 2025 ACM Conference on Applied Artificial Intelligence (AAAI Industry Track). – 2025. – P. 88-97.
7. GenAI Adoption 2024: The Challenge with Enterprise Data // K2view Technical Report. – 2024.