УДК 004.8

ENTERPRISE-RAG-СИСТЕМА ДЛЯ ФИНАНСОВОГО ДОМЕНА БАНКА: ОЦЕНКА ВКЛАДА RETRIEVAL-СЛОЯ И LLM В КАЧЕСТВО ОТВЕТОВ И БИЗНЕС-МЕТРИКИ

Кузнецов Е. А.[1], Потоцкий С. С.[2]
Научный руководитель – доцент Горгадзе А. А.[1]
[1]Университет ИТМО
[2]СПбГЭТУ «ЛЭТИ»
egor03.kuznetsov@yandex.ru

Введение

Крупные языковые модели все активнее используют retrieval-augmented generation для работы с внешними источниками знаний [2]. Вертикальные домены нуждаются в таком подходе особенно сильно, поскольку модели лишены достаточной специализированной экспертизы. Финансовая сфера обзавелась собственными бенчмарками для RAG-систем, проверяющими численный reasoning при анализе отчетности и текстовых массивов [5]. Фреймворк RAGAS автоматически оценивает подобные системы по метрикам фактической достоверности, релевантности извлеченного контекста, соответствия ответа исходному запросу [1]. Архитектура retrieval-слоя и способ встраивания контекста в prompt влияют на итоговое качество ответов, что подтверждают современные обзоры [2]. Банковский enterprise-сценарий делает эту зависимость критичной. Некорректное извлечение информации или ошибки генерации напрямую искажают финансовые расчеты и затрагивают бизнес-метрики [6]. Исследование направлено на создание enterprise-RAG-системы для финансового домена банка, ее оценку и количественное разграничение вклада параметрических знаний языковой модели и retrieval-компонента в качество формируемых ответов [4].\

Основная часть

Методологический подход строится на фреймворке RAGAS, который автоматически измеряет faithfulness, context relevance и answer relevance без необходимости готовить полноценные эталонные ответы [1]. Диагностику retrieval и generation по отдельности обеспечивает ARES: LLM-судьи, обученные на синтетических тройках «запрос – пассаж – ответ», формируют метрики с доверительными интервалами по трем осям качества RAG. Архитектурные решения и конфигурации конвейера retrieval-plus-generation подбираются с опорой на финансовый бенчмарк OmniEval, систематизирующий схемы индексирования и структурирующий многозадачные финансовые сценарии [6].

Численный reasoning реализован через постановку FinQA: многошаговые арифметические операции выполняются над табличной и текстовой отчетностью, результаты сопровождаются программной трассировкой [5]. Разрабатываемая RAG-платформа объединяет векторное хранилище Qdrant, LLM-модель, модуль web-поиска и набор enterprise-критериев, сформированных по мотивам CRUD-RAG [7].

Эффективность RAG-режима сопоставляется с тремя базовыми линиями: LLM-only, retrieval-only и lexical-поиск. Фиксируются метрики RAGAS/ARES и доля корректных FinQA-подобных расчетов. Такая схема эксперимента позволяет явно разграничить вклады архитектуры RAG, качества retrieval-слоя и параметрических знаний LLM, выявляя классы задач, где enterprise-RAG дает статистически значимое преимущество [3].

Выводы

Предложенный подход соединяет принципы автоматизированной оценки RAG-систем с методиками финансовых QA-бенчмарков. Адаптация выполнена специально для enterprise-RAG в банковской сфере [7]. Разработанная методология позволяет аргументированно обосновать преимущества RAG-архитектуры перед альтернативными решениями: изолированным усилением языковых моделей или отдельным совершенствованием поисковых механизмов. Практическая ценность работы выходит за рамки банковского сектора. Результаты применимы к иным высокорисковым доменам, где критически важны точные численные вычисления и строго контролируемое взаимодействие с профессиональной документацией [6].

Литература

1. Es S., James J., Paul L., Nagulapalli M. и др. RAGAS: Automated Evaluation of Retrieval-Augmented Generation [Электронный ресурс] // arXiv preprint arXiv:2309.15217. – 2023. – URL: https://arxiv.org/abs/2309.15217 (дата обращения: 28.02.2026).
2. Gao Y., Xiong Y., Gao X. и др. Retrieval-Augmented Generation for Large Language Models: A Survey [Электронный ресурс] // IEEE Transactions on Knowledge and Data Engineering. – 2024. – URL: https://ieeexplore.ieee.org/document/10468268 (дата обращения: 28.02.2026).
3. Chen J., Han X., Lin H., Sun L. Benchmarking Large Language Models in Retrieval-Augmented Generation [Электронный ресурс] // Proceedings of the AAAI Conference on Artificial Intelligence. – 2024. – Т. 38, № 16. – С. 17795–17803. – URL: https://ojs.aaai.org/index.php/AAAI/article/view/29728 (дата обращения: 28.02.2026).
4. Saad-Falcon J., Khattab O., Potts C., Zaharia M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems [Электронный ресурс] // arXiv preprint arXiv:2311.09476. – 2023. – URL: https://arxiv.org/abs/2311.09476 (дата обращения: 28.02.2026).
5. Chen Z., Chen W., Smiley C. и др. FinQA: A Dataset of Numerical Reasoning over Financial Data [Электронный ресурс] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2021. – С. 3697–3711. – URL: https://aclanthology.org/2021.emnlp-main.300/ (дата обращения: 28.02.2026).
6. Wang S., Tan J., Dou Z., Wen J.-R. OmniEval: An Omnidirectional and Automatic RAG Evaluation Benchmark in Financial Domain [Электронный ресурс] // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. – Suzhou, China: Association for Computational Linguistics, 2025. – С. 5726–5751. – DOI: 10.18653/v1/2025.emnlp-main.292. – URL: https://aclanthology.org/2025.emnlp-main.292/ (дата обращения: 28.02.2026).
7. Lyu Y., Li X., Zhang N. и др. CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models [Электронный ресурс] // arXiv preprint arXiv:2401.17043. – 2024. – URL: https://arxiv.org/abs/2401.17043 (дата обращения: 28.02.2026).