

**Оптимизация Retrieval-Augmented Generation для работы с технической документацией
промышленных компаний
Крайновских В.И. (ИТМО)**

**Научный руководитель – ассистент факультета технологий искусственного интеллекта
Лаушкина А.А.
(ИТМО)**

Введение. Современные промышленные компании ежедневно обрабатывают огромные объемы технической документации, включая спецификации, чертежи, инструкции по эксплуатации и регламенты. Устаревшие методы поиска информации в таких документах часто оказываются недостаточно эффективными, что приводит к увеличению времени обработки данных и повышению вероятности ошибок. Для решения этой проблемы перспективным направлением является использование *Retrieval-Augmented Generation* (RAG), который объединяет информационный поиск и генеративные модели для точного извлечения и генерации ответов [1].

Основная часть. Оптимизация Retrieval-Augmented Generation (RAG) для работы с технической документацией промышленных компаний осуществляется посредством комплексного подхода, включающего усовершенствование эмбединговых моделей, внедрение эффективных стратегий ранжирования, а также использование гибридных методов поиска [2]. В рамках исследования была проведена оценка производительности различных моделей эмбединга, включая BGE-M3, Multilingual-E5 и RuBERT-Tiny. После этого модель BGE-M3 была дообучена с использованием Triplet Loss. Этот подход позволил повысить точность извлечения релевантной информации [3]. Для повышения эффективности поиска реализована двухэтапная стратегия ранжирования retrieved-данных, включающая bi-encoder для первичной фильтрации релевантных документов и cross-encoder для их уточненного реранжирования [4]. Дополнительно разработаны гибридные методы поиска, интегрирующие sparse- и dense-векторные представления, что позволило улучшить релевантность извлекаемых данных [5]. Оценка качества работы системы осуществлялась с использованием метрик качества поиска и генерации, включая Precision@k, Recall@k, Mean Reciprocal Rank (MRR), а также специализированные фреймворки RAGAS и ARES, ориентированные на анализ релевантности и достоверности ответов [6].

Результаты. Экспериментальные исследования показали, что оптимизированная система значительно повысила эффективность поиска и генерации ответов в технической документации промышленных компаний. В частности, дообучение эмбедингера BGE-M3 позволило увеличить точность поиска релевантной информации, что отразилось на росте метрики MRR на 15% по сравнению с базовым подходом [2]. Реализация двухэтапной стратегии ранжирования retrieved-данных, включающей использование bi-encoder и cross-encoder, обеспечила существенное улучшение релевантности найденных документов. Применение гибридных методов поиска, интегрирующих sparse- и dense-векторные представления, привело к сокращению количества нерелевантных результатов, что, в свою очередь, способствовало снижению ошибок, связанных с устаревшей или некорректной информацией [5]. Оценка качества работы системы с использованием метрик Precision@k, Recall@k, а также специализированных подходов RAGAS и ARES подтвердила значительное повышение точности и полноты поиска [6].

Выводы. Результаты исследования подтверждают эффективность предложенных методов оптимизации RAG для обработки технической документации промышленных компаний. Разработанная система позволяет сократить время поиска информации, повысить точность извлекаемых данных и улучшить работу специалистов, занимающихся техническим обслуживанием и анализом документации.

Список использованных источников:

1. Sawarkar K., Mangal A., Solanki S. Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers // 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR). – 2024. – С. 155-161.
2. Shen J., He H., Shen W., Shen T. Enhancing the RAG Retrieval Engine Through Multi-Encoder Fusion // 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). – 2024. – С. 227-230.
3. Lin H., Zhan S., Su J., Zheng H.-T., Wang H. IRSC: A Zero-shot Evaluation Benchmark for Information Retrieval through Semantic Comprehension in Retrieval-Augmented Generation Scenarios // ArXiv. – 2024.
4. Arora D., Kini A., Chowdhury S. R., Natarajan N., Sinha G., Sharma A. GAR-meets-RAG Paradigm for Zero-Shot Information Retrieval // ArXiv. – 2023.
5. Soman S., Roychowdhury S. Observations on Building RAG Systems for Technical Documents // ArXiv. – 2024.
6. Myers S., Miller T. A., Gao Y., Churpek M., Mayampurath A., Dligach D., Afshar M. Lessons Learned on Information Retrieval in Electronic Health Records: A Comparison of Embedding Models and Pooling Strategies // ArXiv. – 2024.