

УДК 004.822

Современные методы генерации, дополненной поиском и их применение в технике и машиностроении

Тронов Кирилл Александрович kirtron@mail.ru
Гагарин Юрий Евгеньевич^(*) gagarin_ye@bmstu.ru

КФ МГТУ им. Н.Э. Баумана, Калуга, Россия

Аннотация. Генерация, дополненная поиском (RAG), — гибридная архитектура, устраняющая ключевые ограничения больших языковых моделей, такие как галлюцинации и зависимость от статических данных, за счет интеграции механизмов информационного поиска с генеративными LLM. В данном обзоре систематизированы последние достижения в области RAG, включая инновационные модели и приложения в специализированных доменах. Целью данного исследования является предоставление исследователям и практикам основополагающего ресурса для понимания потенциала RAG и его направлений в технике и машиностроении.

Ключевые слова: генерация дополненная поиском, информационный поиск, обработка естественного языка, искусственный интеллект, машинное обучение, большие языковые модели

Modern methods of retrieval augmented generation and their application in engineering and mechanical engineering

Tronov Kirill Aleksandrovich kirtron@mail.ru
Gagarin Yuriy Yevgen'yevich^(*) gagarin_ye@bmstu.ru

KB BMSTU, Kaluga, Russia

Abstract. Retrieval-Augmented Generation (RAG) is a hybrid architecture that addresses key limitations of large language models (LLMs), such as hallucinations and dependence on static data, by integrating information retrieval mechanisms with generative LLMs. This review systematizes recent advances in RAG, including innovative models and applications in specialized domains (medicine, programming). The goal of this study is to provide researchers and practitioners with a foundational resource for understanding the potential of RAG and its direction in the field of natural language processing.

Keywords: retrieval-augmented generation, information retrieval, natural language processing, artificial intelligence, machine learning, large language model

Введение. Генерация, дополненная поиском (RAG), — это гибридная архитектура в области обработки естественного языка, которая объединяет механизмы извлечения информации с генеративными большими языковыми моделями. Основная цель RAG — **повысить точность и надежность выходных данных БЯМ**, решая ключевые ограничения традиционных генеративных моделей, таких как галлюцинации и зависимость от статических

данных обучения. Интегрируя знания из внешних баз данных, RAG значительно повышает точность и надежность сгенерированных ответов. Это также обеспечивает происхождение или доказательство того, как был сгенерирован ответ. Кроме того, RAG позволяет постоянно обновлять знания и интегрировать информацию, специфичную для предметной области, что делает БЯМ более актуальными и применимыми в динамичных сценариях.

Методы и материалы. В данном обзоре использован систематический подход к анализу литературы по генерации, дополненной поиском (RAG), с акцентом на последние достижения 2024–2025 гг. Методы включают поиск и отбор публикаций, извлечение данных и их синтез. Поиск проводился в академических базах данных, таких как arXiv, Google Scholar, IEEE Xplore и ACM Digital Library, с использованием ключевых слов: retrieval-augmented generation (RAG), RAG in AI, ontology-based RAG, multimodal RAG, RAG in engineering и RAG applications in manufacturing. Период охвата — с января 2024 г. по сентябрь 2025 г., чтобы захватить актуальные тенденции. Первоначальный отбор включал удаление дубликатов и проверку заголовков/аннотаций на релевантность; всего рассмотрено более 50 публикаций, из которых отобрано 9 ключевых работ для детального анализа.

Результаты. Бурное развитие генерации, дополненной поиском (RAG) в 2024–2025 гг., характеризуется переходом от базовых архитектур к специализированным, адаптивным и мультимодальным системам. Анализ ключевых публикаций выявил три доминирующих направления, отражающих эволюцию задач.

Преодоление фундаментальных ограничений БЯМ (галлюцинации, статичность знаний) в высокоответственных доменах:

- MMed-RAG разработана для борьбы с фактическими галлюцинациями и ненадежностью медицинских больших визуально-языковых моделей [1];
- ICA-RAG решает проблему избыточного и ненужного поиска информации в RAG-системах для клинической диагностики [2].

Интеграция сложных структур знаний (онтологии, графы, таблицы), ранее недоступных для обработки стандартным RAG:

- OG-RAG позволяет большим языковым моделям (БЯМ) адаптироваться к точным, основанным на фактах доменным данным [3].

KnowRA решает проблему отсутствия полного спектра рассуждений (распознавание паттернов, корелляция, логические и причинно-следственные связи) в существующих методах извлечения связей на уровне документов (Doc-RE) [4]:

- RAKG улучшает автоматическое построение графов знаний на уровне документов (KGC), решая проблемы разрешения неоднозначности сущностей и жесткого определения схем [5];

- TableRAG решает ключевые ограничения существующих RAG-подходов при работе с гетерогенными документами, содержащими текст и таблицы [6].

Расширение функциональности БЯМ для решения комплексных задач через координацию инструментов и модальностей:

– CodeRAG решает проблему генерации кода на уровне репозитория для существующих LLM, ограниченных контекстным окном и пониманием длинного кода в сложных средах с зависимостями и доменными знаниями, используя биграф-архитектуру [7];

– HM-RAG устраняет «изоляцию модальностей» с потерей информации, предлагая иерархическую многоагентную архитектуру для динамического синтеза знаний [8];

– RAG-MCP решает проблему раздувания промпта и сложности выбора в БЯМ с растущим числом внешних инструментов (например, по протоколу MCP) [9].

Эти тренды демонстрируют сдвиг от простого «поиска-и-генерации» к когнитивным RAG-системам, где извлечение контекста неразрывно связано с семантическим выравниванием, рассуждениями и динамической адаптацией. Ниже представлена детализация инноваций по каждому направлению.

Обсуждение полученных результатов. Рассмотренные тенденции в развитии RAG — от преодоления ограничений БЯМ до интеграции сложных структур знаний и расширения мультимодальных возможностей находят широкое применение в технике и машиностроении, где точность, эффективность и интеграция доменных знаний критически важны. В частности, в аддитивном производстве (AM), проектировании и контроле качества RAG позволяет автоматизировать сложные задачи, такие как оптимизация структур, выявление аномалий и диагностика дефектов, опираясь на внешние базы знаний без полного переобучения моделей. Одним из ключевых примеров является онтология-ориентированный RAG для анализа данных в AM, где онтология AMDA преобразуется в граф знаний для извлечения релевантных сущностей и генерации контекстно-ориентированных ответов. В проектировании RAG-фреймворк с компонентами Retrieve-Analyze-Generate ускоряет итерации, интегрируя физически-информированные нейронные операторы для симуляций геометрий и многоуровневую кросс-валидацию для оптимизации ресурсов.

Заключение. Таким образом, общая тенденция заключается в переходе от базовых RAG-систем к более умным, адаптивным, специализированным и сложным мультимодальным архитектурам RAG, которые могут не только извлекать релевантную информацию, но и эффективно интегрировать, фильтровать, выравнивать и использовать ее для повышения точности, надежности и способности к рассуждению БЯМ в условиях реального мира и специализированных доменов. С добавлением приложений в машиностроении RAG позиционируется как инструмент математического моделирования для технических задач.

Список источников

- [1] Xia P., Zhu K., Li H., Shi W., Wang S., Zhang L., Zou J., Yao H. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. ArXiv.2410.13085, 2024. <https://doi.org/10.48550/arXiv.2410.13085>

- [2] Jia M., Jia Z., Duan J., Song Y., Wang J. ICA-RAG: Information Completeness Guided Adaptive Retrieval-Augmented Generation for Disease Diagnosis. *ArXiv. 2502.14614*, 2025. <https://doi.org/10.48550/arXiv.2502.14614>
- [3] Sharma K., Kumar P., Li Y. OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models. *ArXiv.2412.15235*, 2024. <https://doi.org/10.48550/arXiv.2412.15235>
- [4] Mai C., Wang Y., Gong Z., Wang H., Huang Y. KnowRA: Knowledge Retrieval Augmented Method for Document-level Relation Extraction with Comprehensive Reasoning Abilities. *ArXiv.2501.00571*, 2024. <https://doi.org/10.48550/arXiv.2501.00571>
- [5] 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.2504.09823*, 2025. <https://doi.org/10.48550/arXiv.2504.09823>
- [6] Yu X., Jian P., Chen C. TableRAG: A Retrieval Augmented Generation Framework for Heterogeneous Document Reasoning. *ArXiv.2506.10380*, 2025. <https://doi.org/10.48550/arXiv.2506.10380>
- [7] Li J., Shi X., Zhang K., Li L., Li G., Tao Z., Liu F., Tao C., Jin Z. CodeRAG: Supportive Code Retrieval on Bigraph for Real-World Code Generation. *ArXiv.2504.10046*, 2025. <https://doi.org/10.48550/arXiv.2504.10046>
- [8] Liu P., Liu X., Yao R., Liu J., Meng S., Wang D., Ma J. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation. *ArXiv.2504.12330*, 2025. <https://doi.org/10.48550/arXiv.2504.12330>
- [9] Gan T., Sun Q. RAG-MCP: Mitigating Prompt Bloat in LLM Tool Selection via Retrieval-Augmented Generation. *ArXiv.2505.03275*, 2025. <https://doi.org/10.48550/arXiv.2505.03275>