УДК 004.8

Бородулин И.В.

магистр технических наук

Омский государственный технический университет

(г. Омск, Россия)

УВЕЛИЧЕНИЕ ТОЧНОСТИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ С ПОМОЩЬЮ РАСШИРЕННОЙ ПОИСКОВОЙ ГЕНЕРАЦИИ

Аннотация: *в последние годы большие языковые модели (LLM) стали ключевым элементом в разработке систем искусственного интеллекта, предназначенных для обработки естественного языка. Однако несмотря на значительные успехи, точность и надежность их ответов остаются предметом для улучшений. В данной статье предлагается методика увеличения точности LLM с использованием технологии расширенной поисковой генерации (RAG), которая позволяет динамически интегрировать информацию из внешних источников данных при генерации ответа. Проведено исследование демонстрирующее, как RAG улучшает качество и релевантность ответов LLM, обеспечивая более точную и информативную генерацию текста.*

Ключевые слова: *искусственный интеллект, обработка естественного языка, большие языковые модели, расширенная поисковая генерация, генерация текста, моделирование знаний, векторное хранилище, поисковые системы.*

Введение. Большие языковые модели (LLM) в последнее время стали революционным шагом в области искусственного интеллекта [1], предоставив возможность генерировать тексты, близкие к человеческим по качеству и смыслу. Тем не менее ограничения в точности и актуальности информации, предоставляемой этими моделями, вызывают необходимость в поиске новых подходов к улучшению их эффективности [2].

Расширенная поисковая генерация. Расширенная поисковая генерация (RAG) представляет собой метод, позволяющий языковым моделям дополнять свои знания [2] за счет поиска и интеграции информации из внешних баз данных в реальном времени. Это позволяет значительно увеличить точность и актуальность генерируемых ответов.

Большие языковые модели, такие как GPT, Gemini, PaLM, Llama, Claude, обучаются на огромных объемах текстовых данных, что позволяет им генерировать текст, имитирующий естественный человеческий язык. Основной недостаток этих моделей заключается в их статичности: они полагаются исключительно на информацию, полученную в ходе первоначального обучения, и не способны обновлять свои знания или адаптироваться к новым данным.

Расширенная поисковая генерация (RAG) предполагает интеграцию LLM с механизмом поиска по специальным векторным базам данных, чтобы обогатить процесс генерации текста актуальной информацией. В основе RAG лежит двухэтапный процесс: сначала выполняется поиск по ключевым словам из запроса пользователя для извлечения релевантной информации, затем полученные данные используются для генерации ответа с помощью LLM. Структурная схема применения RAG вместе с LLM представлена на рисунке 1.

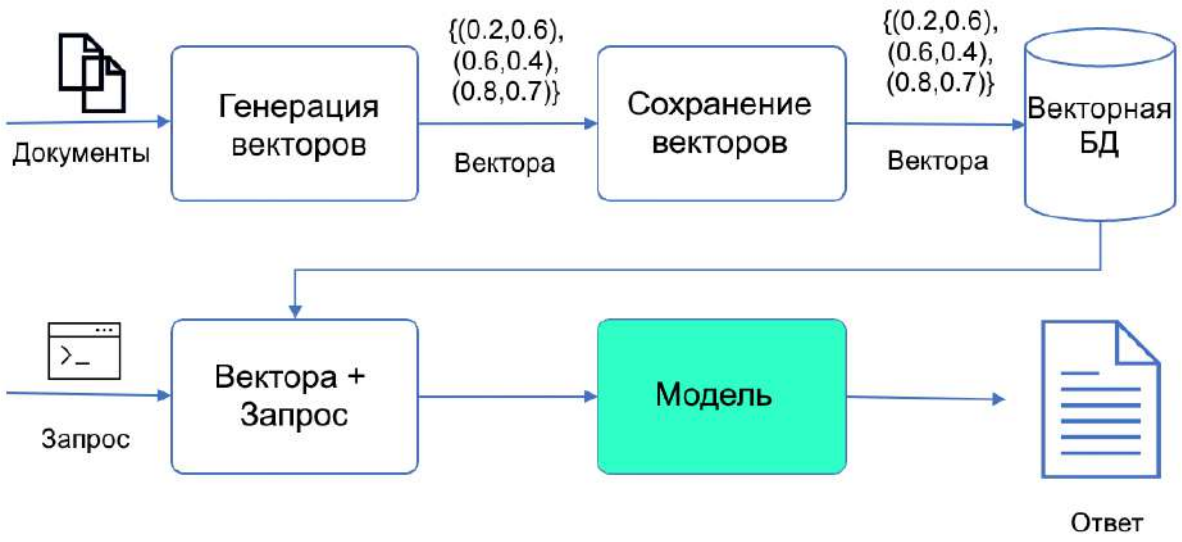

Рис. 1. Процесс генерации текста с помощью RAG.

Для реализации механизма RAG критически важно эффективно интегрировать процесс поиска внешних данных с генерацией текста языковой моделью. Основным инструментом для этого являются векторные представления текста (Embeddings), которые позволяют оценить семантическую близость между запросом пользователя и информацией в базе данных [3].

Процесс генерации ответа на запрос пользователя с использованием векторных представлений происходит следующим образом:

1. Для каждого запроса пользователя создается векторное представление с использованием предварительно обученной языковой модели (например, BERT или GPT). Данные вектора отражают семантическое содержание запроса.
2. Аналогичным образом создаются векторные представления для каждого документа в базе данных, которая может использоваться для поиска дополнительной информации.
3. С помощью различных метрик [4] рассчитывается сходство между embeddings запроса и embeddings в базе данных. Это позволяет определить наиболее релевантные фрагменты данных к запросу пользователя.
4. Фрагменты данных упорядочиваются на основе их сходства с запросом. Высокоранжированные данные считаются наиболее релевантными и выбираются для последующей генерации ответа.

Таким образом, выбранные на основе векторов фрагменты данных используются как дополнительный контекст для языковой модели при генерации ответа. Это позволяет модели генерировать ответы, которые не только семантически связаны с запросом пользователя, но и обогащены актуальной информацией из внешних источников.

Исследование. Для демонстрации влияния технологии RAG на точность LLM было проведено сравнительное исследование. Целью исследования было

выявить различия в точности ответов, генерируемых моделью с использованием и без использования RAG, на примере задачи ответов на вопросы.

Исследование базировалось на двух сценариях:

1. LLM генерировала ответы на вопросы исключительно на основе своих внутренних знаний, полученных в ходе обучения.
2. Перед генерацией ответов LLM интегрировалась с RAG для поиска и интеграции актуальной информации из внешних баз данных.

В качестве модели для исследования была использована GPT-3 и набор вопросов, требующих актуальной информации для ответа по нескольким категориям, например, медицинские данные, актуальные новости, научные открытия [5, 6, 7, 8].

Результаты исследования сравнения точности ответов модели по различным категориям вопросов представлены в таблице 1.

Таблица 1. Сравнение достоверности ответов модели.

Категория вопроса	Точность без RAG	Точность с RAG	Улучшение
Актуальные события	0,50	0,80	+30%
Исторические факты	0,80	0,90	+10%
Научные открытия	0,60	0,85	+25%
Финансовые новости	0,55	0,90	+35%
Географические данные	0,70	0,85	+15%
Культура	0,65	0,80	+15%
Технологии	0,58	0,83	+25%
Медицина и здоровье	0,62	0,88	+26%

Анализ результатов. Результаты исследования показали значительное улучшение в точности ответов при использовании RAG по всем категориям вопросов. Особенно заметное улучшение наблюдается в категориях, требующих актуальной информации, таких как "Актуальные события" и "Финансовые новости", где улучшение составило 30% и 35% соответственно. Также стоит

отметить, что даже в категориях, где предполагалось, что LLM могут обладать достаточным количеством внутренних знаний, таких как "Исторические факты", использование RAG привело к улучшению точности ответов на 10%. Это указывает на то, что даже в областях с относительно стабильной информацией доступ к актуальным данным может улучшить качество ответов.

Заключение. Исследование подтверждает, что использование расширенной поисковой генерации значительно повышает точность ответов больших языковых моделей на вопросы, за счет обеспечения доступа к самой актуальной и точной информации, которая может отсутствовать во внутренних обучающих данных. Это особенно важно для ответов на вопросы, требующих знаний о недавних событиях или данных, которые часто обновляются. Технология RAG открывает новые возможности для улучшения качества и надежности систем искусственного интеллекта, основанных на обработке естественного языка.

СПИСОК ЛИТЕРАТУРЫ:

1. A Survey of Large Language Models // arXiv.org : сайт. – URL: https://arxiv.org/abs/2303.18223 (дата обращения: 02.03.2024);
2. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection // arXiv.org : сайт. – URL: https://arxiv.org/abs/2310.11511 (дата обращения: 03.03.2024);
3. Language-agnostic BERT Sentence Embedding // arXiv.org : сайт. - URL: https://arxiv.org/abs/2007.01852 (дата обращения: 06.03.2024);
4. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018;
5. Wikipedia, the free encyclopedia : сайт. – URL: https://en.wikipedia.org/ (дата обращения: 10.03.2024);
6. Our World in Data : сайт. - URL: https://ourworldindata.org/ (дата обращения: 10.03.2024);

7. Google Dataset Search : сайт. - URL: https://datasetsearch.research.google.com/ (дата обращения: 12.03.2024);
8. GeoNames : сайт. - URL: http://www.geonames.org/ (дата обращения: 12.03.2024)

Borodulin I.V.
Omsk State Technical University
(Omsk, Russia)

IMPROVING ACCURACY OF LARGE LANGUAGE MODELS USING RETRIEVAL AUGMENTED GENERATION

***Abstract**: in recent years, Large Language Models (LLMs) have become a key element in the development of artificial intelligence systems designed for natural language processing. However, despite significant advances, the accuracy and reliability of their responses remain subject to improvement. This paper proposes a methodology to improve the accuracy of LLMs by utilizing a technique called Retrieval Augmented Generation (RAG) that dynamically integrates information from external data sources during response generation. A research is conducted to demonstrate how RAG improves the quality and relevance of LLM responses by providing more accurate and informative text generation.*

***Keywords**: artificial intelligence, natural language processing, large language models, retrieval augmented generation, text generation, knowledge modeling, vector database, search engines.*