Разработка адаптивного RAG подхода для создания эффективной вопросно-ответной системы

Антропова Е.Г.
antropova.eg@phystech.edu

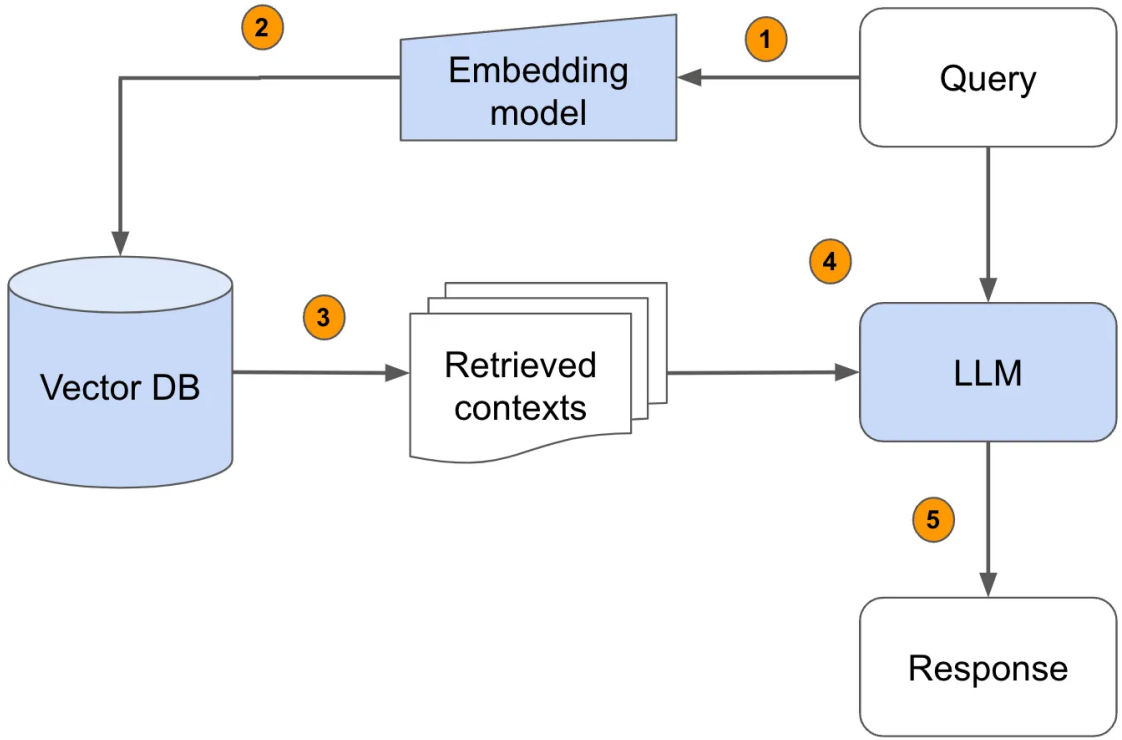
Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар

Retrieval Augmented Generation (RAG)

Подход, который объединяет два этапа:

1. Поиск релевантной информации (retrieval) – модель извлекает нужные данные из базы знаний.
2. Генерация ответа (generation) – языковая модель использует найденные данные для формулирования осмысленного ответа.



Retrieval Augmented Generation (RAG)

RAG сталкивается с такими проблемами, как высокие вычислительные затраты на поиск и генерацию ответов.

Цель работы

Разработка адаптивного подхода Retrieval-Augmented Generation (RAG), направленного на улучшение работы QA-системы.

Формальная постановка задачи

Пусть

- Q — множество текстовых пользовательских запросов,
- D — множество документов (база знаний),
- $\mathcal{R}(q, D)$ — функция извлечения релевантных документов по запросу $q \in Q$, возвращающая подмножество документов $D_q \subseteq D$.
- $\mathcal{G}(q, D_q)$ — генеративная функция, которая на основе $q \in Q$ и найденных документов D_q формирует ответ $\mathbf{a}$.

Цель работы: Разработать функции извлечения документов и генерации ответа $\mathcal{R}^*$ и $\mathcal{G}^*$, улучшающие качество извлечения документов и генерации ответов.

$$D_q = \mathcal{R}^*(q, D) \text{ and } \mathbf{a} = \mathcal{G}^*(q, D_q).$$

QA-система для табличных данных

Проблемы:

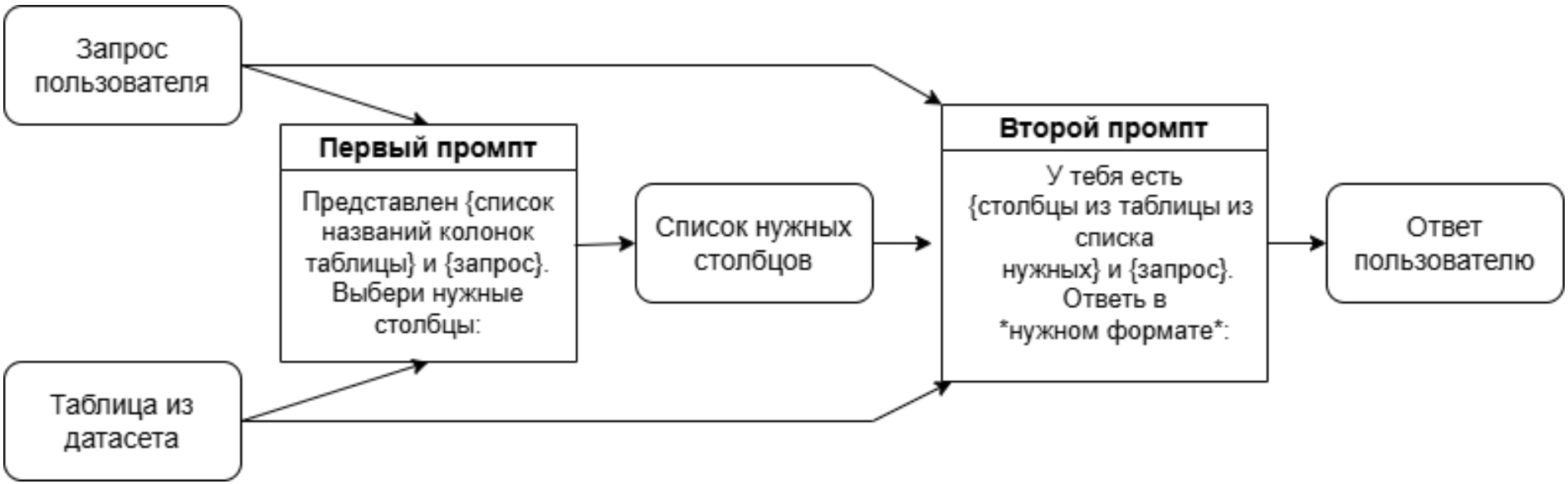
- Неочевидность способа поиска релевантной информации по таблице;
- Понижение качества ответа при увеличении размера таблицы;
- Невозможность обработки больших таблиц с помощью LLM.

Предлагается улучшить QA-систему, подавая в LLM не всю таблицу, а её часть.
Например, подход[1] извлекает сегменты таблицы на основе запроса.
Но не всегда нужен/применим поиск по строкам.

Chen et al. (2024) «TableRAG: Million-Token Table Understanding with Language Models»

Column Augmented Generation (CAG)

RAG-подход для обработки табличных данных. LLM из всех названий столбцов выбирает те, которые могут ответить на вопрос пользователя, далее ответ генерируется на основе выбранных столбцов.



Эксперименты

Подход CAG опробован на датасете DataBench[2] соревнования[3]. Представлены результаты выбора столбцов на трёх LLM. Результаты сравниваются с эталонными, представленными организаторами.

Model	Precision	Recall	F1
Llama-3.3-70B-Instruct	84.60	98.33	90.95
Llama-3.1-8B-Instruct	61.80	96.90	75.46
DeepSeek-R1-32B	77.28	99.01	86.81

В данном датасете таблицы не содержат названия строк, достаточно поиска по столбцам.

Grijalba et al. (2024). «Question Answering over Tabular Data with DataBench: A Large-Scale Empirical Evaluation of LLMs»

Os'es Grijalba et al. (2025). SemEval-2025 task 8: Question answering over tabular data. In Proceedings of the 19th International Workshop on Semantic Evalua- tion (SemEval-2025)

Эксперименты

Представлены окончательные результаты применения CAG.
Применение подхода значительно повысило значения метрик.

Model	EM
Llama-3.1-8B-Instruct	27.78
+CAG	37.93
Llama-3.3-70B-Instruct	49.81
+CAG	64.37
DeepSeek-R1-32B	77.78
+CAG	83.52