Проект посвящён разработке и исследованию RAG-системы (Retrieval-Augmented Generation) на русском языке с использованием различных LLM и векторных баз данных. Целью авторок стало сравнение качества генерации ответов на вопросы в базовой модели и в улучшенной версии с retrieval-частью, а также поиск оптимальных параметров и архитектурных решений.

В работе реализованы как retrieval-, так и non-retrieval варианты системы. Участицы также сравнили разные модели генерации (RuadaptQwen и Vikhr-Gemma). В проекте учтена специфика работы с русскоязычным текстом (использованы адаптированные модели, применена лемматизация и стемминг).

Некоторые аспекты презентации показались нам не очень прозрачными, что немного затруднило для нас интерпретацию выводов:

- В разделе экспериментов упоминаются параметры window_size, top_key, num_beams, max_new_tokens. Для нас остались непонятными параметры window_size и top_key, поэтому было бы полезно включить краткие определения.
- Авторки используют разбиение текста на чанки. Было бы интересно узнать, почему именно такой размер чанка использовался, а также рассмотреть влияние размера чанка на качество извлечения и генерации, поскольку значение параметра window_size в исследовании не варьируется.
- Интересной идеей является повтор релевантного чанка в промпте, но в презентации не показано, как именно это реализуется и влияет ли это на метрику. Один-два примера вывода модели с пояснением были бы полезны для понимания.
- Осталось неясным, по какому принципу подбирались конкретные комбинации параметров. Кроме того, было бы интересно увидеть, как работают модели большего размера.
- Используются разные значения параметров для retrieval и non-retrieval вариантов, что делает прямое сравнение затруднительным. В идеале стоило бы выровнять параметры или пояснить, почему такое сравнение корректно.
- Мы советуем пару улучшений визуализации некоторых слайдов:
    - На слайде 11 не указано, какая именно модель использовалась для получения векторов.
    - Слайды 13 и 14 трудно интерпретировать без подробного объяснения осей и переменных. На слайде 14 есть отрицательные и положительные значения — мы не смогли понять, что они означают, какие параметры при этом менялись и был ли это retrieval-вариант.

В целом участницы проекта проделали значительный объём работы, успешно используя релевантные инструменты. Любопытно, что при использовании retrieval качество генерации не улучшилось, а лишь стало стабильнее.