Отчет по заданию №3 «RAG»

Выполнила: Учар Айгуль, 324 группа

Для сравнения были взяты модели HuggingFaceTB/SmolLM2-135M-Instruct и TinyLlama/TinyLlama-1.1B-Chat-v1.0

Программная реализация:

- Производится удаление дублирующихся вопросов.
- Удаляются дубликаты чанков с помощью множества seen_documents.
- В функции Qdrant.from_documents создается векторная база данных в памяти, где каждый документ индексируется с использованием эмбеддингов, созданных при помощи HuggingFaceEmbedding.
- Настройка моделей и RAG системы: в функции RetrievalQA.from_chain_type создаются цепочки RAG для генерации ответов на основе контекста из Qdrant.
- В функции generate_answers генерируются ответы на топ-10 вопросов. Ответ извлекается из выданного текста с помощью функции extract_answer.
- С помощью метрик ROUGE-L и BERTScore оценивается качество ответов.

Значения метрик:

	Модель	ROUGE-L	BERT score
1	HuggingFaceTB/SmolLM2-135M-Instruct	0.000	0.528
2	TinyLlama/TinyLlama-1.1B-Chat-v1.0	0.365	0.802

TinyLlama/TinyLlama-1.1B-Chat-v1.0 (1.1 миллиарда параметров) показала значительно более высокие результаты по сравнению с HuggingFaceTB/SmolLM2-135M-Instruct (135 миллионов параметров), что логично: TinyLlama имеет почти в 10 раз больше параметров, что позволяет ей лучше улавливать сложные зависимости в данных и генерировать более точные и семантически близкие ответы.

Ответы 1 модели часто не соответствуют вопросу и содержат бессвязный или повторяющийся текст.

ROUGE-L оценивает точное совпадение последовательностей, поэтому нулевой результат указывает на полное отсутствие соответствия ответов эталону.

BERT score анализирует близость сгенерированного и эталонного ответов через эмбеддинги BERT, учитывая контекст. Даже если ответы модели содержат ключевые слова или частично пересекаются по смыслу с эталоном (например, упоминание "команд" или "Суперкубков"), BERT score фиксирует эту слабую связь, что объясняет более высокий балл (0.528).

Ответы 2 модели более точные и соответствуют вопросу. Например, на вопрос о количестве очков, уступленных защитой Пэнтерс, модель кратко и правильно отвечает "308 очков".

Вывод:

Сравнение моделей HuggingFaceTB/SmolLM2-135M-Instruct (135 млн параметров) и TinyLlama/TinyLlama-1.1B-Chat-v1.0 (1.1 млрд параметров) показывает, что увеличение размера модели напрямую влияет на качество генерации.