Audio-lecture helper

LLM with RAG

github

Над проектом работали:
Марьин Геннадий ФКН ПМИ 3 курс
Тяжова Наталья ФКН ПМИ 3 курс

Цель и задачи

Цель: по входному аудиофайлу выдать его конспект или его более короткую версию, а также предоставить функционал ответа на вопросы по данному входному файлу, используя встроенную базу данных

Задачи:

- Транскрибирование аудиофайла
- Суммаризация транскрипции и поиск ответов на ее основе
- Создание внутренней БД
- Тестирование полученных результатов
- Разработка телеграм бота для визуализации

Актуальность: разработанная нейронная сеть может значительно упростить жизнь студентов, сэкономив время на написании конспектов по лекциям, и ответив на оставшиеся вопросы по теме

Используемые технологии и идеи

Whisper: нейронная сеть от OpenAI для решения задачи ASR (Automatic Speech Recognition)

Llama3: LLM, SoTa в мире открытых больших языковых моделей

RAG: метод для расширения знаний LLM с помощью подмешивания дополнительной информации

Основные библиотеки:

Langchain, transformers, telebot...

Дополнительные идеи - Semantic Segmentation and Sequential Chain: из-за того, что лекция длится 80 минут, ее транскрипция составляет около 85k символов, поэтому их невозможно все передать сразу в LLM; чтобы решить данную проблему, использовалась идея семантического разбиения текста, полученные чанки последовательно передавались в LLM

Анализ существующего решения:

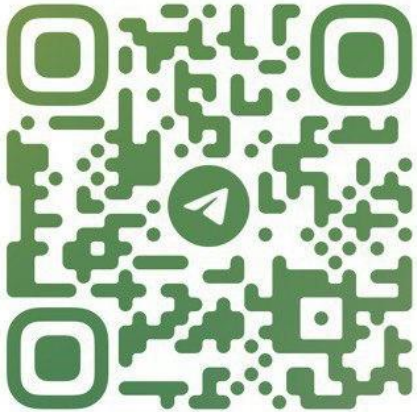
Будем оценивать работу нашей модели по следующим характеристикам:

1. Точность ответа/суммаризации
2. Полнота ответа/суммаризации
3. Соответствие формату
4. Согласованность с БД

С помощью экспериментов удалось показать:

1. Использование RAG помогло модели отвечать на поставленные по лекции вопросы довольно конкретно (пример: short_audio/1.png)
2. Разбиение текста на чанки значительно повысило полноту ответа (с ~1500 до ~30000 символов: long_audio/summary_short vs long_audio/summary_tuned.pdf)
3. Настройка промптов помогла улучшить соответствие формату лекционных записей (long_audio/summary.pdf vs long_audio/summary_tuned.pdf)
4. Благодаря настроенным промптам модель не отвечает на вопросы, не связанные с БД (short_audio/coordination.png)

Спасибо за внимание!



@FCS_VK_BOT