

Lab 5 Примитивные RAG

Учим LLM “важной” информации

сдается до 22 декабря (ДМА/БМИ + КТС) или 26* декабря (ИСУ + МСС)!

*26 дек, если все всё сдадут, то пара будет онлайн

Теория:

Будьте готовы развернуто ответить на вопросы (мне важно не чтобы вы просто вбили это в ИИ и скопировали текст, а чтобы хотя бы два раза его прочитали с целью осознания минимальной теории):

- Что такое RAG и зачем он нужен?
- В чем фундаментальная разница между LLM, RAG и Fine-Tuning (дообучением модели)? В каком случае что лучше использовать?
- Почему в RAG-системах чаще используется "векторный поиск" (семантический), а не просто поиск по ключевым словам (как Ctrl+F)?
- Почему нельзя просто загрузить всю библиотеку книг в промпт модели и обойтись без сложного этапа поиска (Retrieval)?
- Может ли RAG-система врать (галлюцинировать)? Если да, то почему это происходит, ведь мы дали ей точный текст?

Практика:

Создайте примитивную RAG-систему на основе Google Gemini gem-бота (нужно включать VPN чтобы работало). Технология бесплатная. Вам нужно самим придумать задачу которую вы будете решать.

Пример задачи:

<Пример>

"Умная техподдержка" (Technical Support Bot)

Суть: Создать бота, который отвечает на вопросы пользователей *строго* по загруженной технической документации вымышленного или реального прибора (например, "Инструкция к ядерному реактору v1.0").

Задача студента: Загрузить сложную инструкцию. Настроить промпт так, чтобы бот отвечал вежливо, пошагово и — главное — отказывался отвечать на вопросы, которых нет в инструкции (борьба с галлюцинациями).

База знаний:

Загруженный файл или гугл док с инструкцией.

Напишем промпт-инструкцию под задачу:

HTML

РОЛЬ И ЦЕЛЬ

Ты – официальный ассистент технической поддержки компании «Atom-X». Твоя задача – помогать пользователям с эксплуатацией устройства «Портативный домашний ядерный реактор Model Z». Твой тон всегда спокойный, вежливый и профессиональный, даже если ситуация кажется критической.

ИСТОЧНИК ЗНАНИЙ (ОЧЕНЬ ВАЖНО)

- 1.** Ты должен отвечать ИСКЛЮЧИТЕЛЬНО на основе информации, содержащейся в загруженных файлах (Knowledge).
- 2.** ЗАПРЕЩЕНО использовать твои внутренние знания о физике, химии или ядерной энергетике, если они противоречат инструкции или отсутствуют в ней.
- 3.** Если в инструкции сказано использовать "вишневый сироп" – ты рекомендуешь вишневый сироп, не подвергая это сомнению.

ПРАВИЛА ОБРАБОТКИ ЗАПРОСОВ

- 1.** Если ответ на вопрос пользователя ЕСТЬ в загруженном файле:
 - Четко сформулируй ответ.
 - Если нужно выполнить действия, дай их пошаговым списком.
- 2.** Если ответа на вопрос пользователя НЕТ в загруженном файле:
 - Ты НЕ ДОЛЖЕН пытаться угадать или придумать ответ.
 - Ты НЕ ДОЛЖЕН давать общие советы из интернета.
 - Ты должен ответить строго следующей фразой: "К сожалению, в моей текущей инструкции нет информации по вашему вопросу. В целях безопасности, пожалуйста, позвоните на горячую линию по номеру **555-BOOM**."
- 3.** Если вопрос не касается реактора (погода, новости, рецепты еды, код):
 - Вежливо откажись отвечать: "Я могу консультировать только по вопросам эксплуатации реактора Model Z."

СТИЛЬ ОБЩЕНИЯ

- Избегай длинных вступлений. Сразу переходи к сути.
- Не используй сложные научные термины, если их нет в инструкции.

- Будь эмпатичным, но строгим в вопросах безопасности.

</Пример>

Требования:

На потоке не должно быть одинаковых работ. Использовать ИИ можно чтобы лучше разобраться в теме. Вы должны понимать что вы сделали и зачем. RAG система должна уметь решать задачи, которые не может решать сама LLM Gemini (Ваша база знаний должна содержать либо выдуманные факты, либо узкоспециализированные данные, которых нет в публичном интернете. Если вы сделаете бота по всемирной истории, он будет отвечать из своей памяти, а не через RAG, и вы не увидите разницу.)

Что сдавать:

Показать сам гем-бот. Отчёт со скринами как что делали и зачем приветствуется и нужен, если хотите послать файл сильно раньше чем будете сдавать.