?????? LLM, RAG

Генеративный ИИ относится к алгоритмам, которые могут генерировать новый контент, в отличие от анализа существующих данных или воздействия на них, как более традиционные системы машинного обучения с прогнозированием или искусственного интеллекта.

Температура модели - степень вариативности (креативности). При t=0.1 креативности нет.

Параметр top-p выдает элементы в зависимости от вероятности (суммы). Может быть 4, ...

Параметр top-k просто например 3 значения с наивысшей вероятностью. Будет только 3 значения.

RAG

В RAG, модель не обучается на внешних данных. Вместо этого:

- Текст из документов разбивается на фрагменты.
- Для каждого фрагмента создается embedding (вектор).
- Векторы сохраняются в базу (ChromaDB, FAISS, Qdrant и т.д.).
- При вопросе пользователя ищутся подходящие фрагменты.
- Они подставляются в промпт модели Ollama.

Например, пользователь спрашивает: "Где хранятся периоды договора?". Система находит кусок документа: "Period_dogar таблица хранения периодов". Отправляет модели примерно такой запрос:

"Контекст:

Period_dogar таблица хранения периодов
Reestr_dogar основная таблица хранения договоров

Вопрос:
Где хранятся периоды договора?"

Модель отвечает: "Периоды договора хранятся в таблице Period_dogar".

Revision #5
Created 6 April 2025 16:03:19 by Admin
Updated 5 June 2026 04:29:04 by Admin