Управление языковыми моделями

Технические требования

Настройка НА кластера - критичный раздел, но сейчас пока не актуально.

Мне хватило следующего ПК:

Процессор	Intel Xeon E5-2670 v3 @2.3GHz (даже не средний □) Во время обработки грузился на 60%.
ОП	Всего 32 Gb На обученной модели во время обработки вопроса в пиках подскакивало только до 17 Gb Просто ollama в фоне - 11 Gb
SSD	Для размещения модели deepseek-r1:7b потребовалось 5 Gb
Видео	Не использовалось, слишком старая. Да, не особо быстро, иногда полного ответа нужно было ждать секунд 30.
ОС	Windows

Для построения векторного индекса по одному файлу word размером 100 страниц потребовалось 35 минут.

Запуск модели

Использовал менеджер моделей Ollama ollama.com Установщик. Затем управление через cmd.

Команда	Описание
ollama run model_name	Скачать, установить и запустить модель `ollama run deepseek-r1:7b`
ollama list	Список установленных моделей

Команда	Описание
ollama rm model_name	Удаление модели

После запуска по умолчанию http://localhost:11434/ запускается API.

Revision #2
Created 5 April 2025 12:18:03 by Admin
Updated 6 April 2025 05:59:11 by Admin