

Тестирование и сравнение нового функционала LM Studio при работе с нейронными сетями, развернутыми локально с онлайн решениями

Звайгзне Алексей Юрьевич

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

Данное исследование посвящено тестированию и сравнению нового функционала LM Studio в контексте работы с нейронными сетями, развернутыми локально. LM Studio — это платформа для разработки и обучения языковых моделей, обладающая широким набором инструментов для работы с данными, моделями и инфраструктурой. Проведен сравнительный анализ работы полученных результатов работы, развернутой локальной нейросети и онлайн решения на примере работы Perplexity.

Ключевые слова: LM Studio, нейронные сети, локальное развертывание, тестирование, производительность, эффективность.

Testing and comparing the new LM Studio functionality when working with neural networks deployed locally with online solutions

Zvaigzne Alexey Yurievich

Sholom-Aleichem Priamursky State University

Student

Abstract

This study is devoted to testing and comparing the new LM Studio functionality in the context of working with locally deployed neural networks. LM Studio is a platform for developing and teaching language models, with a wide range of tools for working with data, models, and infrastructure.

A comparative analysis of the work of the obtained results of the deployed local neural network and the online solution is carried out using the example of Perplexity.

Keywords: LM Studio, neural networks, local deployment, testing, performance, efficiency.

1 Введение

1.1 Актуальность

В настоящее время наблюдается стремительный рост популярности искусственного интеллекта и, в частности, глубокого обучения. Нейронные сети находят широкое применение в различных областях, от обработки естественного языка до компьютерного зрения. Локальное развертывание нейронных сетей приобретает всё большую актуальность, поскольку:

Обеспечивает конфиденциальность и безопасность данных, данные остаются на локальном устройстве, исключая риски утечки или несанкционированного доступа.

Уменьшает зависимость от интернета, работа нейронной сети не зависит от стабильности интернет-соединения.

Позволяет оптимизировать время реакции, отсутствие необходимости передавать данные в облако сокращает время обработки запросов.

Новый функционал LM Studio, позволяет загружать различные документы для работы с их содержимым позволяя в более быстром порядке обучить нейронную сеть по заданному материалу.

Данное исследование может иметь практическую значимость, так как позволяет оценить эффективность и выгоду использования LM Studio при локальном развертывании нейронных сетей в сравнении с онлайн решениями.

1.2 Обзор исследований

В статье Д.А. Павлов описал потенциальный вред использования искусственных нейросетей [1]. И. И. Багаев произвел анализ перехода внимания научного сообщества от основного средства для распознавания образов при помощи методов машинного обучения без применения нейронных сетей к сверточным нейронным сетям как к наилучшему способу для классификации объектов на изображениях [2]. А.М.Плотникова производит анализ влияния распространения технологий искусственного интеллекта, в частности нейронных сетей, вызывает активное обсуждение их влияния на медиа и общество [3]. В дискуссионной статье Т. В. Черниговская концентрирует внимание на вопросах человечности нейронных сетей сравнивая компьютерную нейронную сеть с человеком [4].

1.3 Цель исследования

Целью исследования является оценка практической ценности нового функционала LM Studio [5] для пользователей, работающих с нейронными сетями в локальной среде в сравнении с функционалом онлайн сервиса Perplexity [6].

2 Материалы и методы

Процесс подключения языковой модели нейронной сети произведен в среде работы с языковыми моделями нейронных сетей LM studio версии 0.3.5 (Build 2), выбраны, установлены и настроены языковая модель Gemma 2 9B Instruct 5.13 GB для сравнения с популярным условно бесплатным онлайн сервисом, поддерживающим работу с файлами и доступом в интернет Perplexity.

Также для быстрой и корректной работы, развернутой локально модели потребуется мощная вычислительная техника, в рамках статьи для работы будет использоваться ноутбук Maibenben с дискретной видеокартой Nvidia RTX 4060 (laptop) с 8 Гб видеопамати, стандарта GDDR6, 16 Гб оперативной

памяти стандарта DDR5, с тактовой частотой 4800 МГц, спецификации процессора не требуются в связи с тем, что для работы модели будет использовать только работа графического видеоускорителя.

Для сравнения работы моделей будет использоваться заранее заготовленный документ.

3 Результаты и обсуждения

3.1 Подготовка к работе локальной модели

Для начала работы скачивается и устанавливается средство работы с языковыми моделями нейронных сетей LM studio с официального сайта (рис. 1) [7]. Данное программное обеспечение является интерфейсом для взаимодействия с моделями позволяет скачивать для работы языковые модели нейронных сетей с ресурса Hugging Face [8] в виде контейнеров формата gguf.

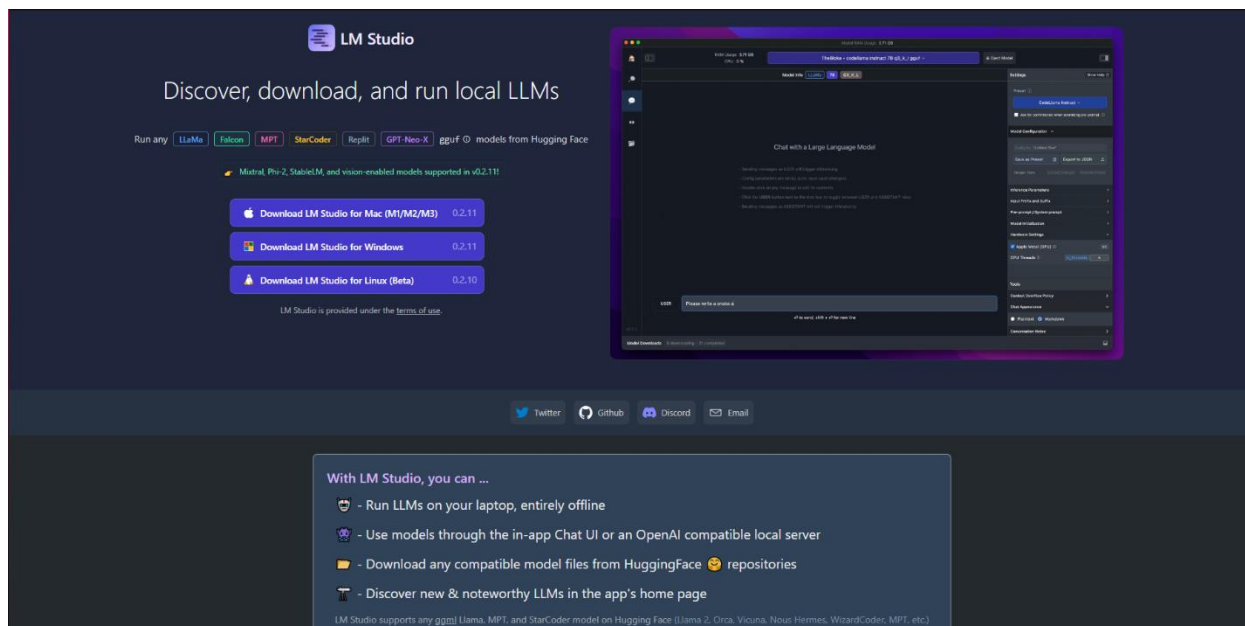


Рисунок 1. Официальный сайт LM studio

Выбрав версию для определенной операционной системы начнется загрузка установочного файла. Загрузив файл производится стандартная процедура установки. После становки программа запускается с помощью ярлыка на рабочем столе. При запуске программы будет отображена вкладка с поиском и установкой предварительно обученных различных типов языковых моделей нейронных сетей от разных авторов размещенных на Hugging face (рис. 2).

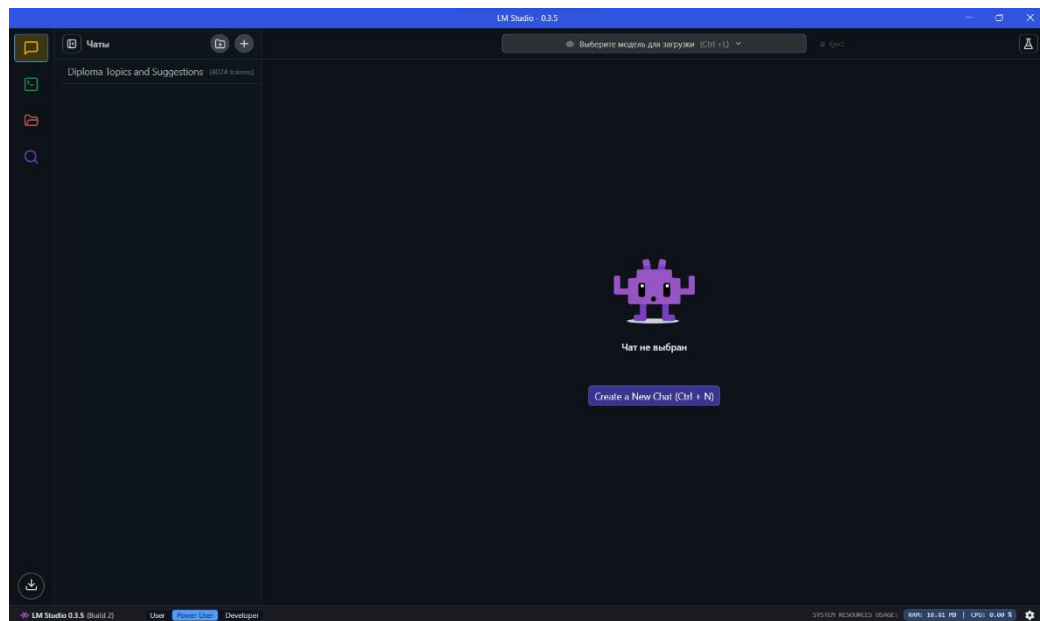


Рисунок 2. Главный экран LM studio

В рамках статьи была выбрана языковая модель нейронных сетей Gemma 2 9B Instruct 5.13 GB (рис. 3) [8] среднего размера семейства моделей Gemma 2 на 9 миллиардов параметров. Созданный компанией Google на основе тех же исследований и технологий, которые использовались при создании моделей Gemini. В данной модели используется Q3_K_L система квантования позволяющая оптимизировать занимаемый объем данных на носителе и требования к видеопамяти, во вред точности работы модели.

Во время работы модели, потребляемая память в простое составляет порядка 6,10 ГБ (рис. 4). Также стоит обратить внимание на то, что если выбран графический ускоритель для работы, то модель будет загружаться в оперативную память видеокарты. При недостаточном объеме оперативной памяти при попытке загрузки модели будет выдаваться ошибка о нехватке памяти (out of memory).

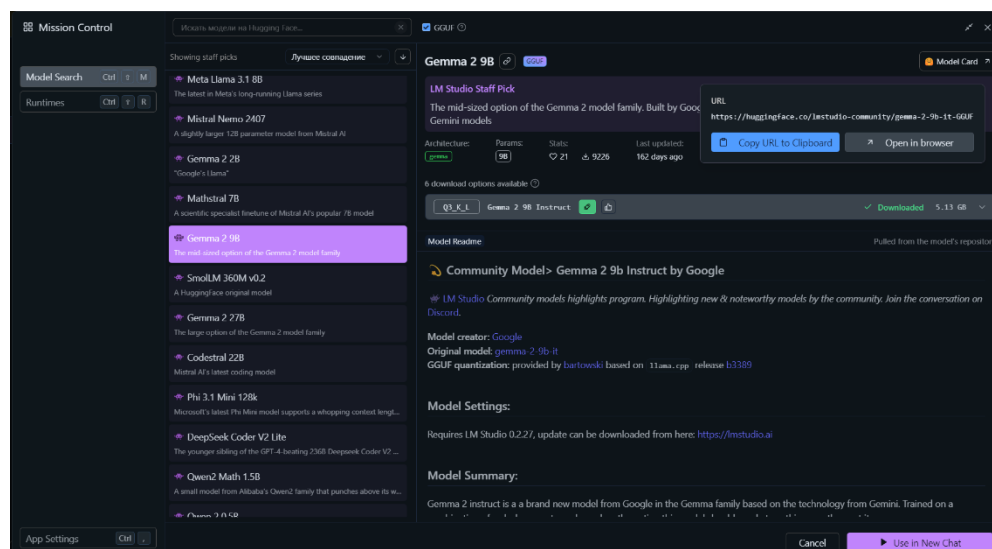


Рисунок 3. Gemma 2 9B доступные языковые модели для загрузки

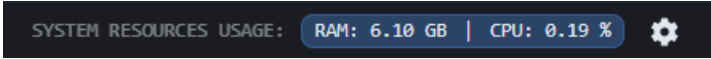


Рисунок 4. Занимаемый объем видеопамати моделью

После установки моделей в левой части экрана выбрать раздел Чат (рис. 5).

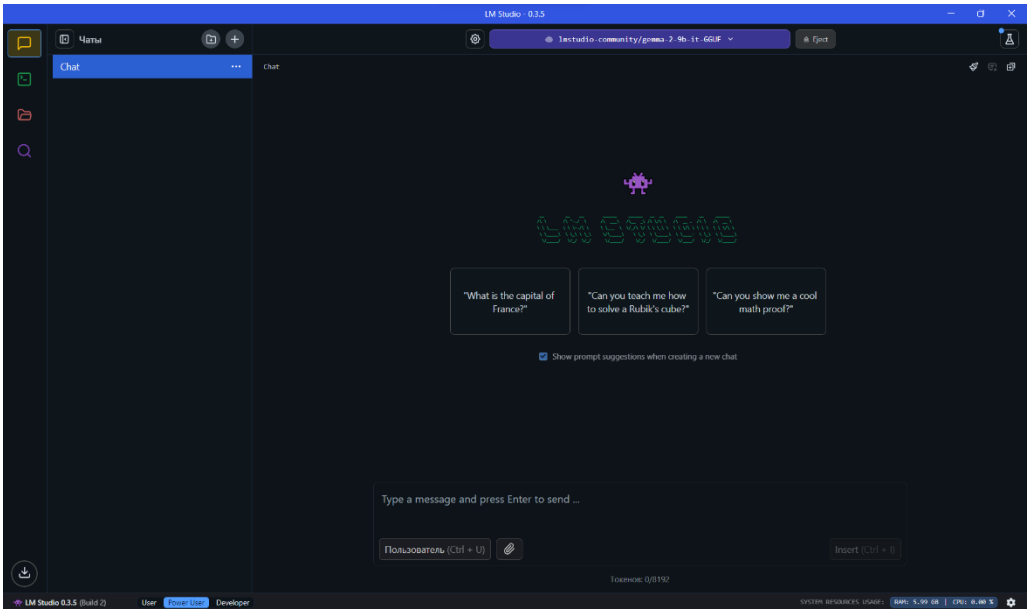


Рисунок 6. Окно чата

3.2 Работа с локально развернутой моделью нейронной сети

В работе с моделью было изменен лишь объем токенов для работы с моделью во избежание перегрузки входных данных для модели в сообщении с 4096 до 8192 токенов на сообщение (рис. 7) после этого необходимо перезагрузить модель.

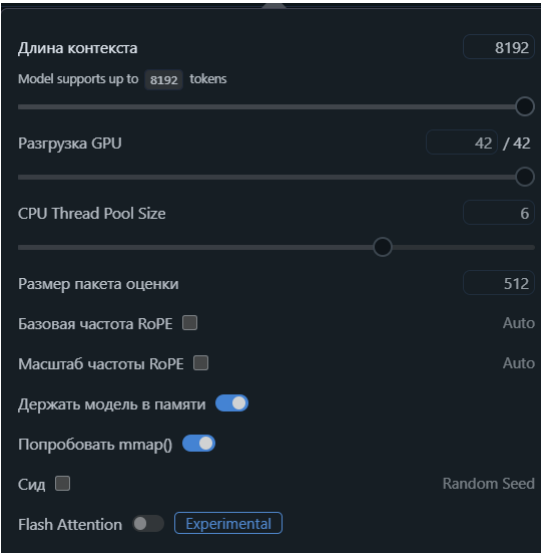


Рисунок 7. Настройка длины контекста входного сообщения

Для проверки работоспособности модели был задан тестовый вопрос (рис. 8).

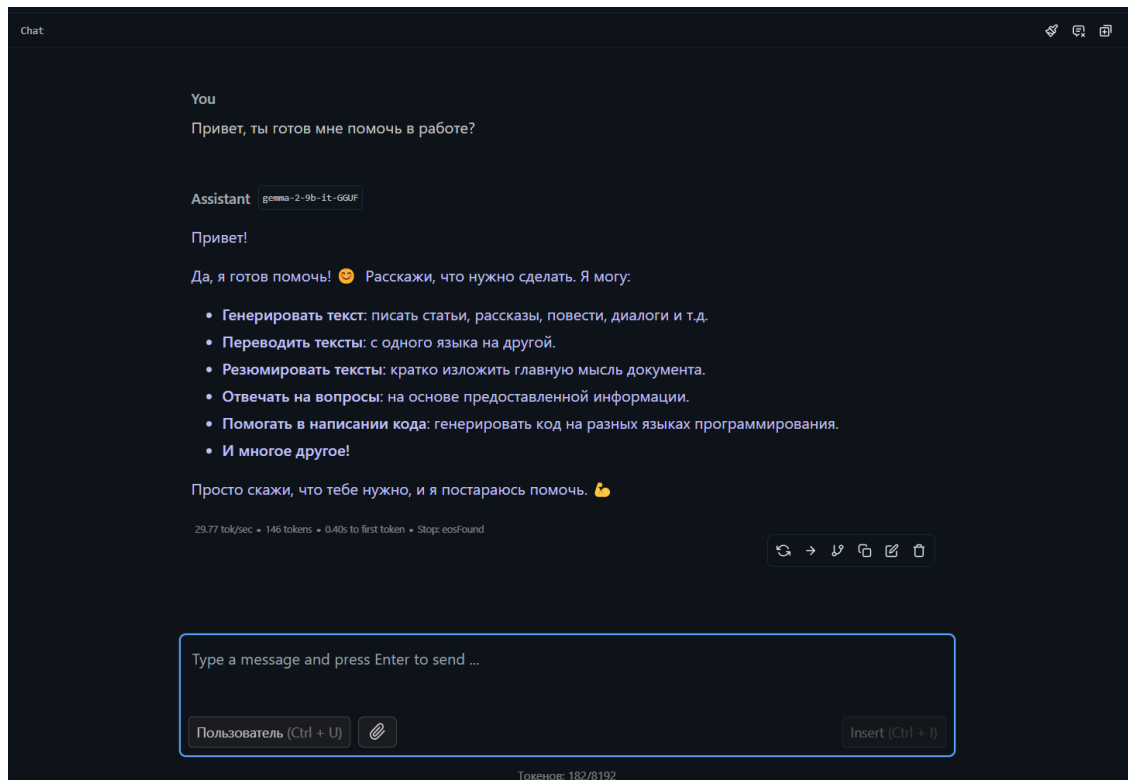


Рисунок 8. Пример работы Gemma

Для сравнения двух разных моделей нейронных сетей будет выбран одинаковый файл формата PDF и задан одинаковый запрос обработки текста.

Так для обеих моделей был выбран файл предыдущей статьи по развертыванию локально нейросети в контексте работы с файлом в интерфейсе LM Studio был создан новый чат [9].

Первично загружен файл, после загрузки, в чате было выдано сообщение на английском языке (рис. 9).

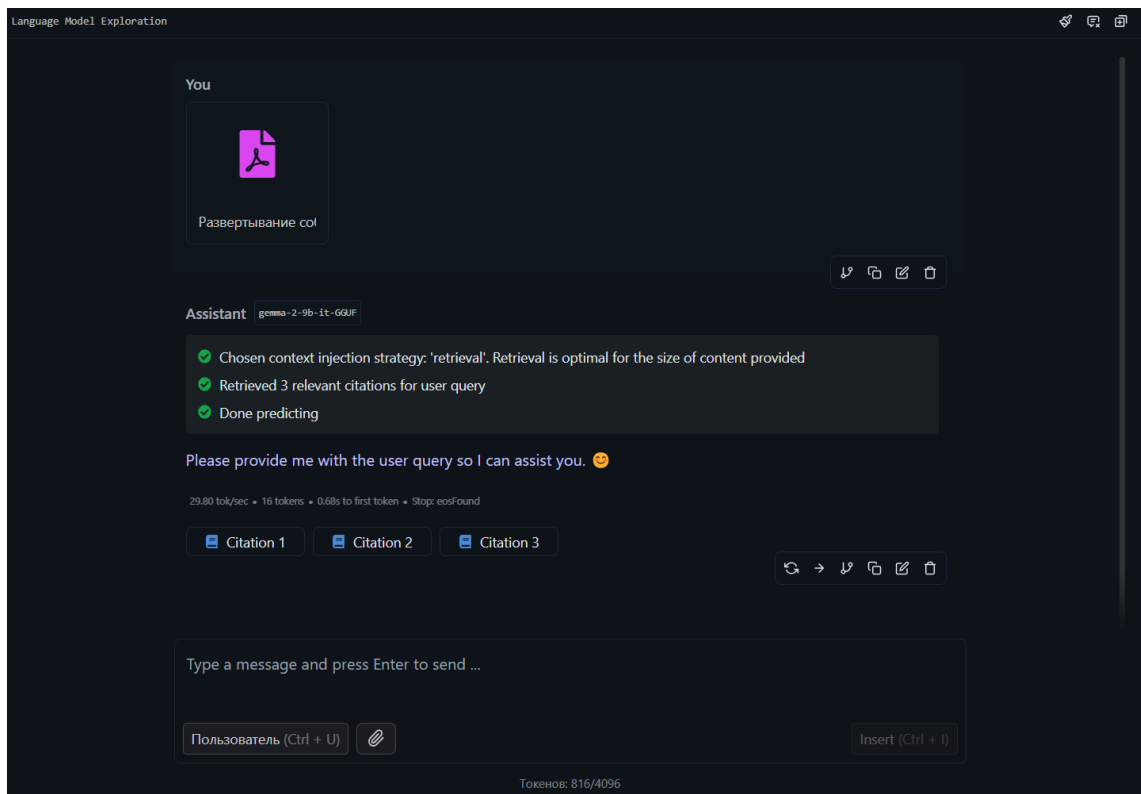


Рисунок 9. Вывод языковой модели после ввода документа в систему

По интерфейсу вывода было определено, что система поделила содержимое текста на цитаты.

При дальнейшей работе был задан вопрос (рис. 10):

«Выведи материалы и методы, описываемые в данной статье»

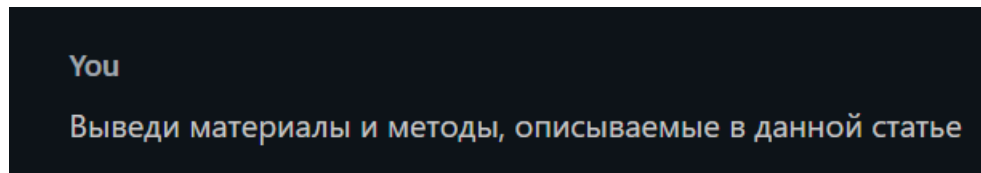


Рисунок 10. Сообщение с запросом к модели

Модель при выводе информации опиралась на ранее составленные цитирования и не смогла дать корректный ответ, также ответ был выдан на английском языке (рис. 11).

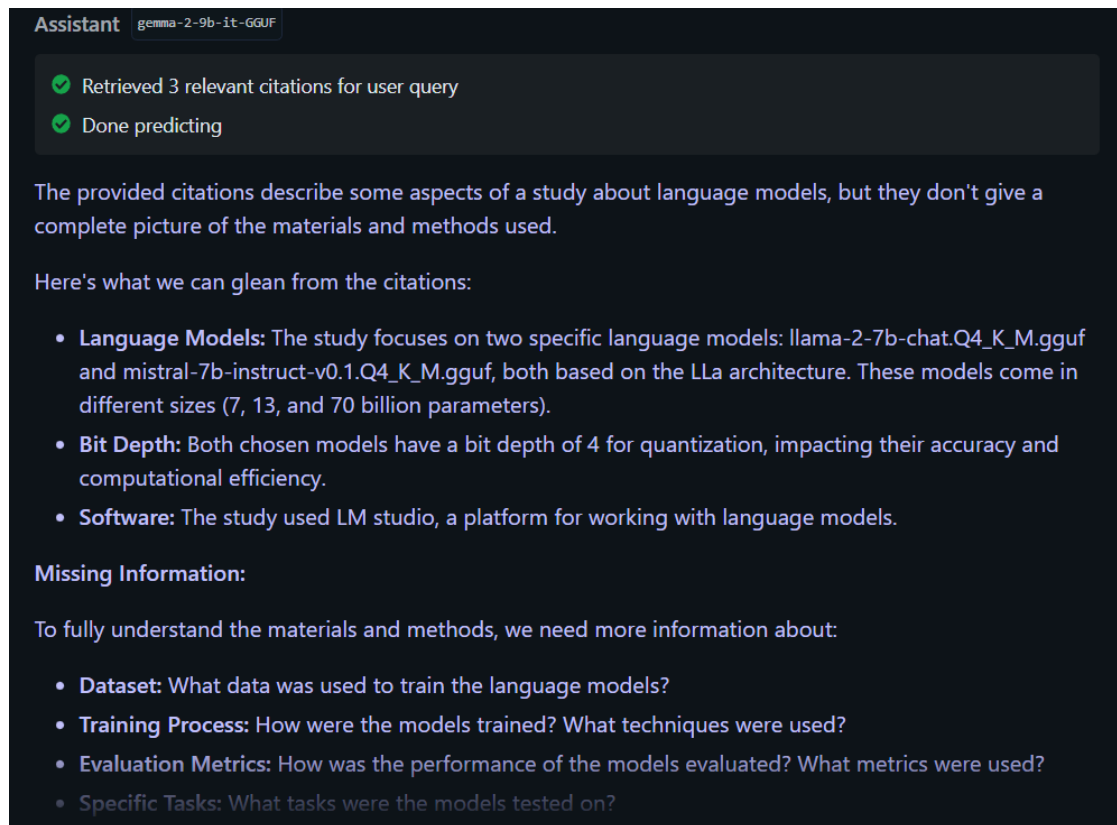


Рисунок 11. Вывод модели по данным цитирования

Отсюда определено, что программа разбивает входной текст на цитирования для оптимизации дальнейшей работы с файлом.

3.3 Повторный тест работы нейронной сети

Для нормализации работы нейронной сети, в расширенной конфигурации модели были внесены правки, а входной массив данных был загружен текстом в поле ввода (рис. 12), также для чистоты эксперимента был создан новый чат.

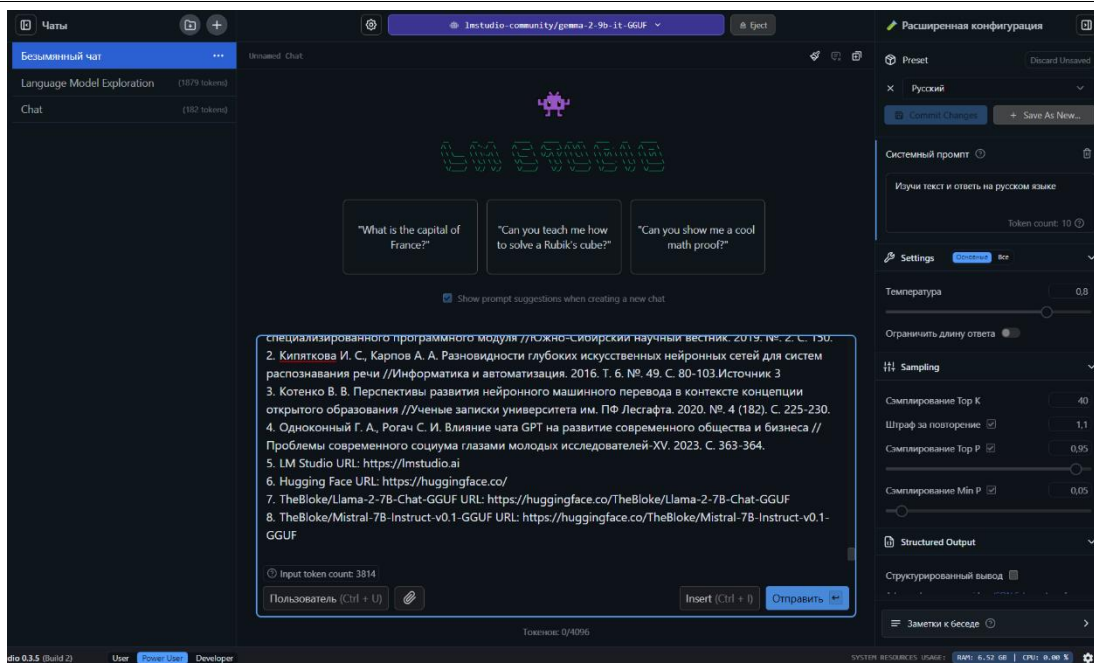


Рисунок 12. Создание нового чата для внесения текста вручную

Нейросеть дала следующий вывод (рис. 13):

Статья описывает процесс развертывания и настройки собственной языковой модели нейронной сети на локальной машине с использованием инструмента LM studio.

Основные моменты:

- **Актуальность:** Авторы подчеркивают важность локального развертывания языковых моделей для обеспечения конфиденциальности, контроля над инфраструктурой и ускорения экспериментальных циклов.
- **Методы:** Статья подробно описывает процесс выбора и установки языковой модели из Hugging Face, настройки параметров и тестирования работы модели в LM studio.
- **Реализация:** Для исследования использовались две модели: LLaMA 2 7B chat и Mistral 7B Instruct v0.1. Q4 K M.
- **Результаты:** Авторы сравнили ответы, полученные от двух моделей на заданный вопрос о создании SQL запросов для ИС автомойки. Оба ответа содержали информацию о создании нужных таблиц, но LLaMA 2 предоставила более подробное описание и структуру.
- **Выводы:** Авторы отмечают, что локальные языковые модели могут быть полезными в частных организациях с ограниченным доступом к сети Интернет.

Общие впечатления:

Статья представляет собой хороший обзор процесса развертывания и использования собственной языковой модели нейронной сети на локальной машине. Автор подробно описал шаги настройки LM studio, выбор моделей, настройку параметров и сравнение результатов работы двух разных моделей.

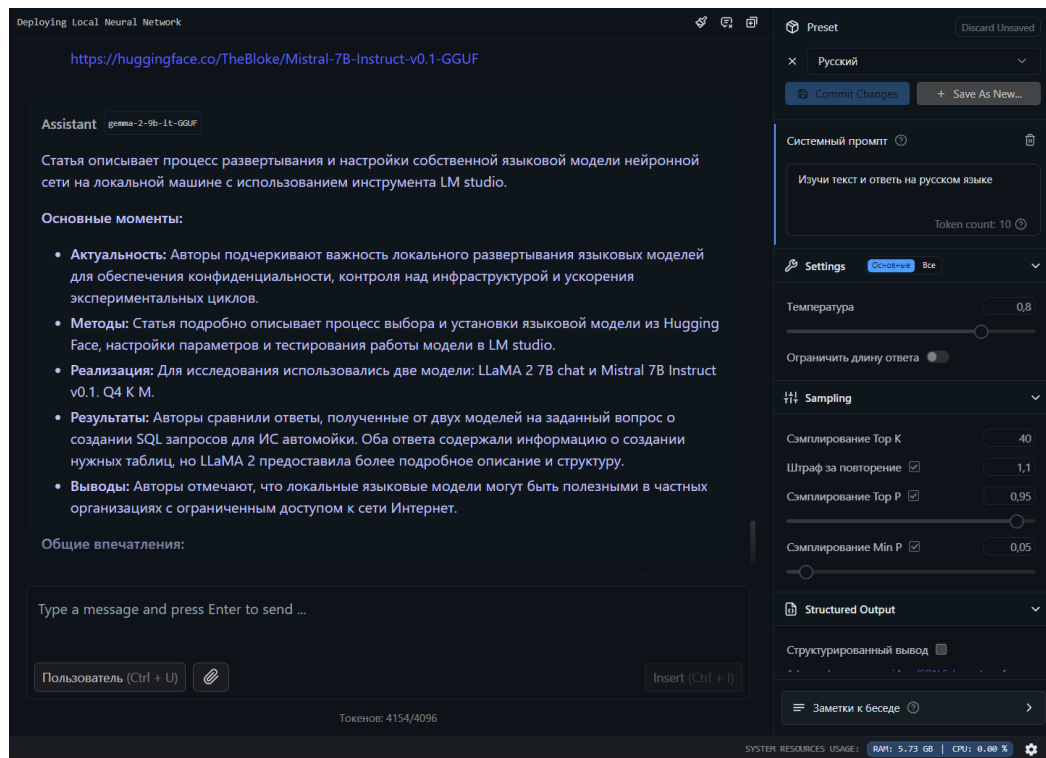


Рисунок 13. Вывод результата

Далее был задан аналогичный вопрос по тексту (рис. 14).

Модель выдала более корректный результат:

Материалы и методы:

1. Инструментальное ПО:

- LM studio: используется для загрузки и настройки языковых моделей нейронных сетей. Скачивается с официального сайта (<https://lmstudio.ai>).

- Hugging Face: используется как источник предобученных языковых моделей, доступных в формате gguf. (<https://huggingface.co/>)

2. Языковые модели:

- Llama-2 7b chat.Q4_K_M: модель с архитектурой LLa, доступна на Hugging Face.

- Mistral 7b instruct v0.1.Q4_K_M: модель с архитектурой LLa, доступна на Hugging Face.

3. Технические требования:

- Мощная вычислительная техника: графический ускоритель Nvidia 3070 и 16 ГБ оперативной памяти для работы с большими моделями.

Выводы из статьи:

- Развертывание локальной версии языковой модели предоставляет ряд преимуществ, таких как контроль над данными, повышенная безопасность и возможность использования специализированных аппаратных ресурсов.

- LM studio - удобный инструмент для работы с различными языковыми моделями нейронных сетей, доступными на Hugging Face.

- Выбор между разными языковыми моделями зависит от конкретных задач и требований к точности и производительности.

Рекомендации:

- Для более полного сравнения моделей было бы полезно провести эксперименты с различными типами запросов и оценить качество ответов каждой модели.

- В статье можно было бы добавить информацию о том, как оптимизировать работу моделей на локальной машине для повышения производительности и снижения потребления ресурсов.

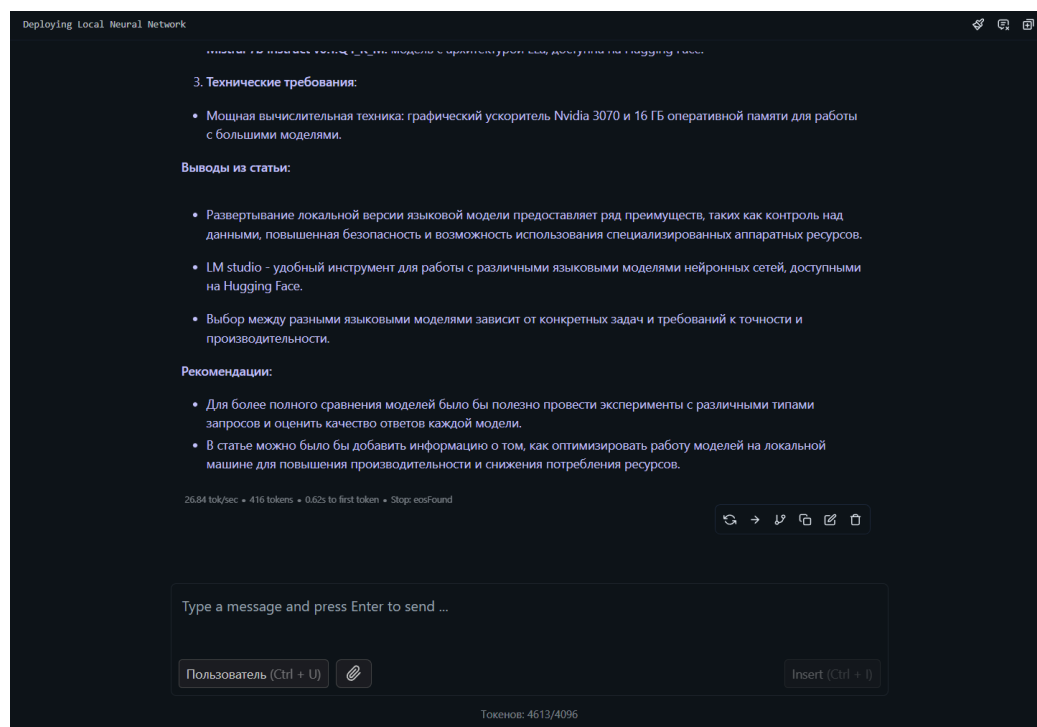


Рисунок 14. Вывод результата работы модели по заданному тексту

Сравнивая два полученных результата, можно сделать промежуточные выводы, что инструмент подгрузки файла в данный момент ещё не совершенен.

3.4 Тестирование работы нейронной сети Perplexity

В связи с тем, что сервис является условно бесплатным, для работы есть ежедневные ограничения на выполнения операций в режиме Pro в количестве – трех запросов в сутки с использованием расширенного функционала (рис. 15).

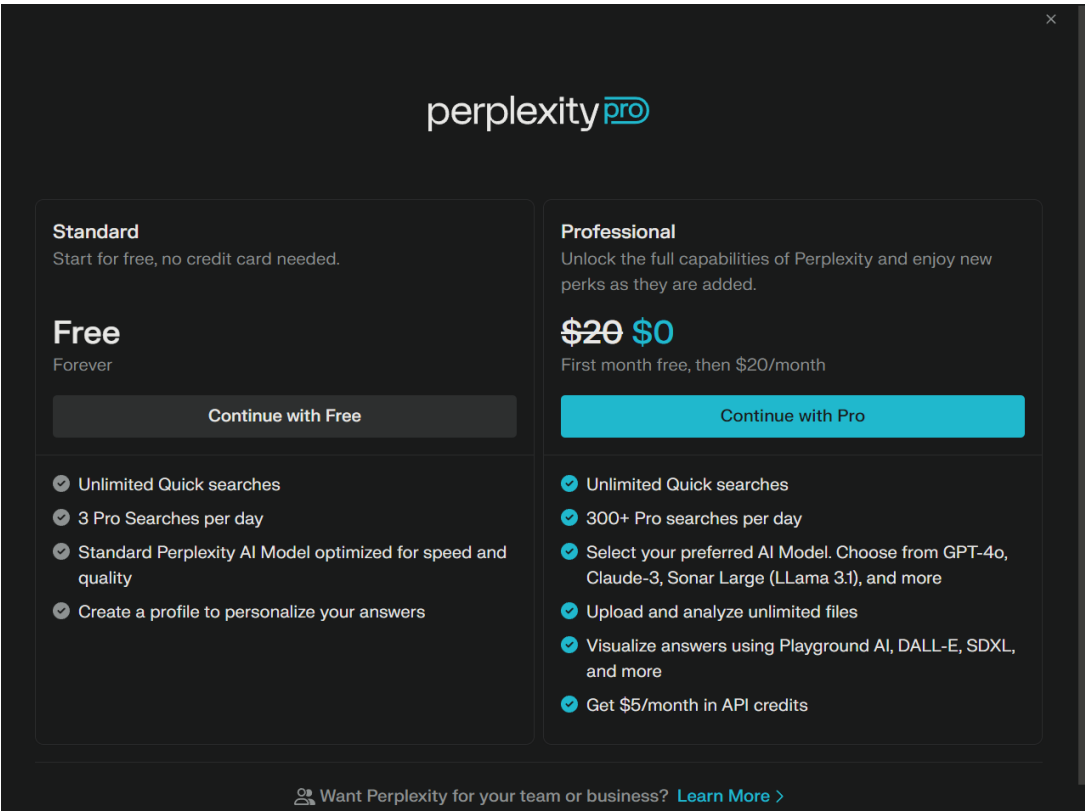


Рисунок 15. Страница выбора тарифа

Для корректной работы системы необходима авторизация. После авторизации можно выполнить три запроса, подгружая необходимые файлы (рис. 16).

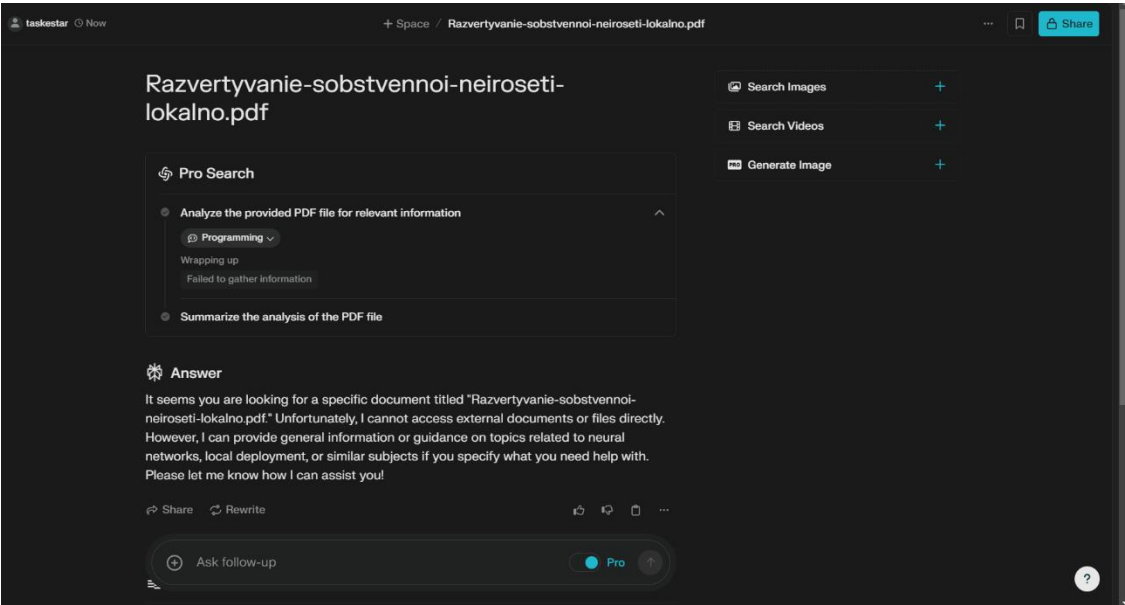


Рисунок 16. Результат загрузки документа

Исходя из первичного вывода документа, можно предположить что нейросеть не смогла обработать запрос, для проверки был введен аналогичный предыдущим тестам вопрос (рис. 17):

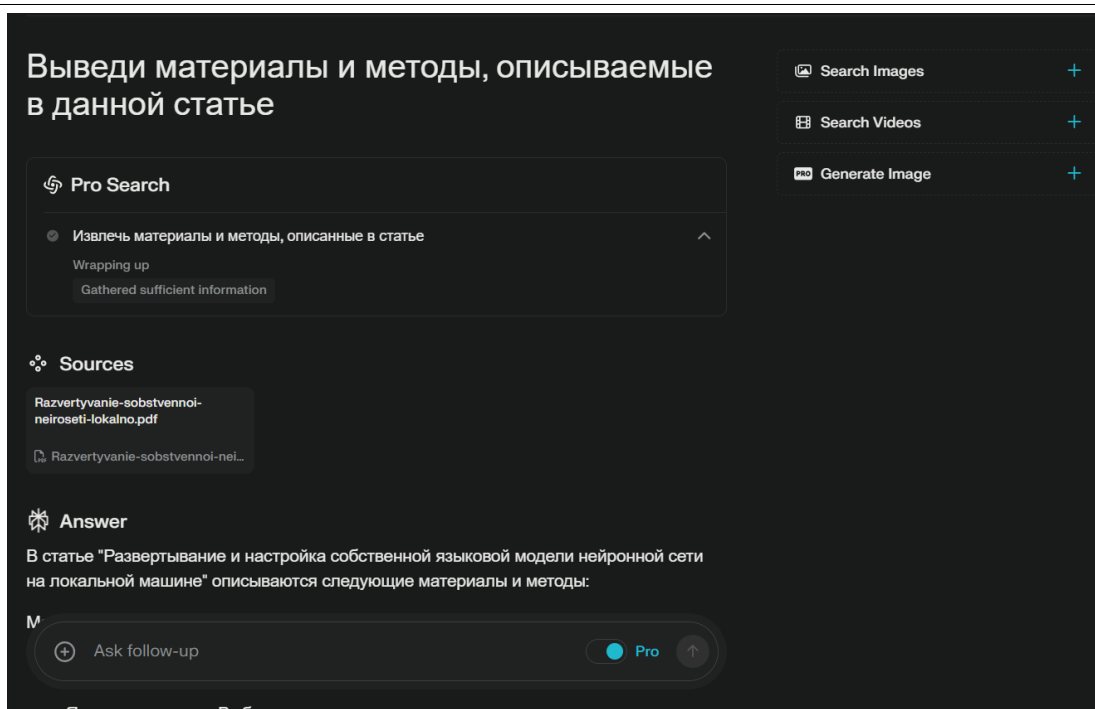


Рисунок 17. Результаты вывода работы Perplexity

При вводе аналогичного запроса опираясь на выданный модели файл, можно определить, что несмотря на первичный отказ в работе с файлом, система может дать необходимый ответ. Возможно данные результаты как-то могут быть связаны с искусственными ограничениями работы нейросети, внедренными разработчиками, ради достижения информационной безопасности при работе.

Результат работы нейронной сети

В статье "Развертывание и настройка собственной языковой модели нейронной сети на локальной машине" описываются следующие материалы и методы.

Материалы:

- Программное обеспечение: Используются инструменты LM Studio для работы с языковыми моделями нейронных сетей.
- Языковые модели: Выбраны и настроены две языковые модели:
- Mistral 7B Instruct v0.1.Q4 K M
- LLAMA 2 7B chat.Q4 K M
- Оборудование: Для работы моделей требуется мощная вычислительная техника, в частности графический ускоритель Nvidia 3070 и 16 ГБ оперативной памяти.

Методы:

1. Установка LM Studio: Сначала скачивается и устанавливается LM Studio с официального сайта, что позволяет загружать языковые модели с платформы Hugging Face в формате контейнеров gguf.
2. Выбор и настройка моделей: После установки пользователю необходимо выбрать нужные языковые модели в интерфейсе LM Studio и

настроить их параметры, включая битность квантования (в данном случае выбрана битность 4 бита).

3. Изменение параметров взаимодействия: Настраиваются параметры, такие как первоначальный запрос (Prompt), подключение графического ускорителя, количество слоев модели и длина контекста токена.

4. Работа с моделями: После настройки моделей производится их перезагрузка, после чего начинается работа через чат, где задаются одинаковые запросы для обеих моделей для сравнения их ответов.

5. Сравнение результатов: Для проверки эффективности моделей задаются одинаковые SQL-запросы на создание таблиц, и результаты анализируются на предмет точности и полноты ответов.

Эти методы позволяют развернуть локальную версию языковой модели нейронной сети и оценить её производительность в сравнении с другими моделями.

3.5 Ввод текста в модель напрямую через диалоговое окно

При попытке вставить аналогичный текст в новый чат система формирует новый текстовый файл (рис. 18).

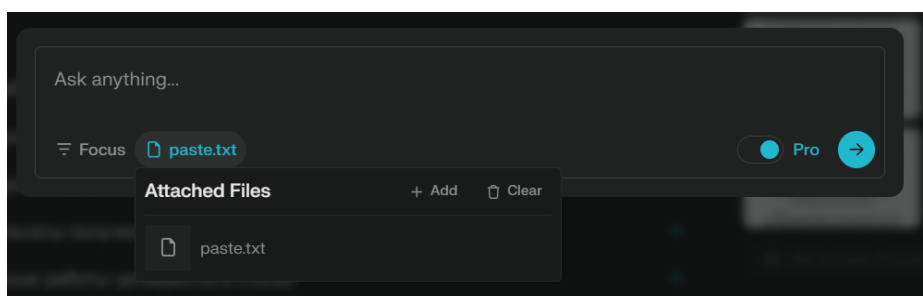


Рисунок 18. Добавления заданного текста в ручную

При загрузке текста напрямую, модель восприняла текст статьи как вопрос и выдало краткое содержание статьи (рис. 19):

Развёртывание и настройка языковой модели нейронной сети на локальной машине

****Аннотация

В данной статье описывается процесс развёртывания локальной версии языковой модели нейронной сети с использованием программного обеспечения LM Studio. Рассматриваются выбор и настройка языковых моделей, а также оценка их производительности.

Введение

Актуальность

Развёртывание индивидуальных языковых моделей на локальных вычислительных ресурсах становится важным направлением в области искусственного интеллекта. Локальное развёртывание обеспечивает:

- Удобство управления и хранения данных

- Полный контроль над инфраструктурой
- Более эффективные эксперименты и обучение моделей на ограниченных данных
- Возможность работы оффлайн
- Экономическую эффективность для небольших исследовательских групп

Цель исследования

Целью данного исследования является развертывание языковой модели на локальной машине с использованием LM Studio, что позволит взаимодействовать с моделью без подключения к Интернету.

Материалы и методы

Процесс подключения языковой модели осуществляется в среде LM Studio, где были выбраны и настроены две модели:

- Mistral 7B Instruct v0.1.Q4 K M
- LLAMA 2 7B chat.Q4 K M

Для работы моделей требуется мощная вычислительная техника, в частности графический ускоритель Nvidia 3070 и 16 ГБ оперативной памяти.

Результаты и обсуждения

Установка LM Studio

1. Скачайте и установите LM Studio с официального сайта.
2. Запустите программу и выберите языковые модели для загрузки из Hugging Face.

Настройка моделей

После установки моделей необходимо настроить параметры в разделе AI chat:

- Выбор модели
- Установка параметров (Prompt, количество слоев, длина контекста токена)

Примеры SQL-запросов

Для сравнения двух моделей были заданы одинаковые запросы. Пример SQL-запроса для создания таблиц для информационной системы автомобиля:

```
sql
-- Создаем таблицу "автоматические_заправки"
CREATE TABLE автоматические_заправки (
  id INT PRIMARY KEY,
  номер_заправки INT,
  кол-во_литров DECIMAL(10, 2),
  стоимость DECIMAL(10, 2)
);
```

Обе модели выдали аналогичные результаты с небольшими различиями в формулировках и детализации.

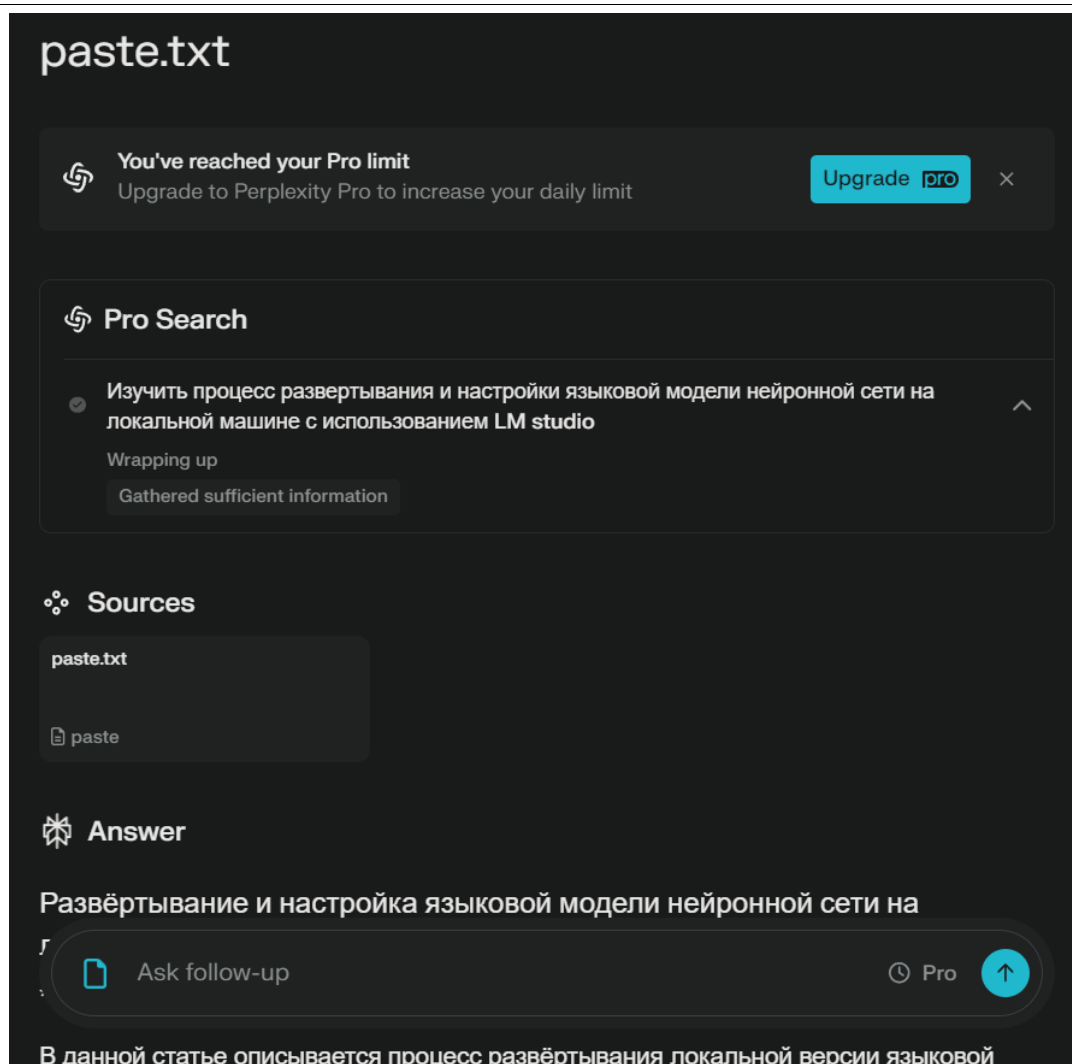


Рисунок 19. Результат работы нейросети при прямой загрузке текста

При попытке ввода аналогичного запроса был получен следующий результат (рис. 20):

Материалы и методы

Процесс подключения языковой модели нейронной сети был осуществлён в среде LM Studio. Для работы были выбраны, установлены и настроены следующие языковые модели:

Mistral 7B Instruct v0.1.Q4 K M

LLAMA 2 7B chat.Q4 K M

Для корректной работы моделей требуется мощная вычислительная техника, в частности:

Графический ускоритель Nvidia 3070

16 ГБ оперативной памяти

Эти компоненты обеспечивают необходимую производительность для запуска и тестирования языковых моделей.

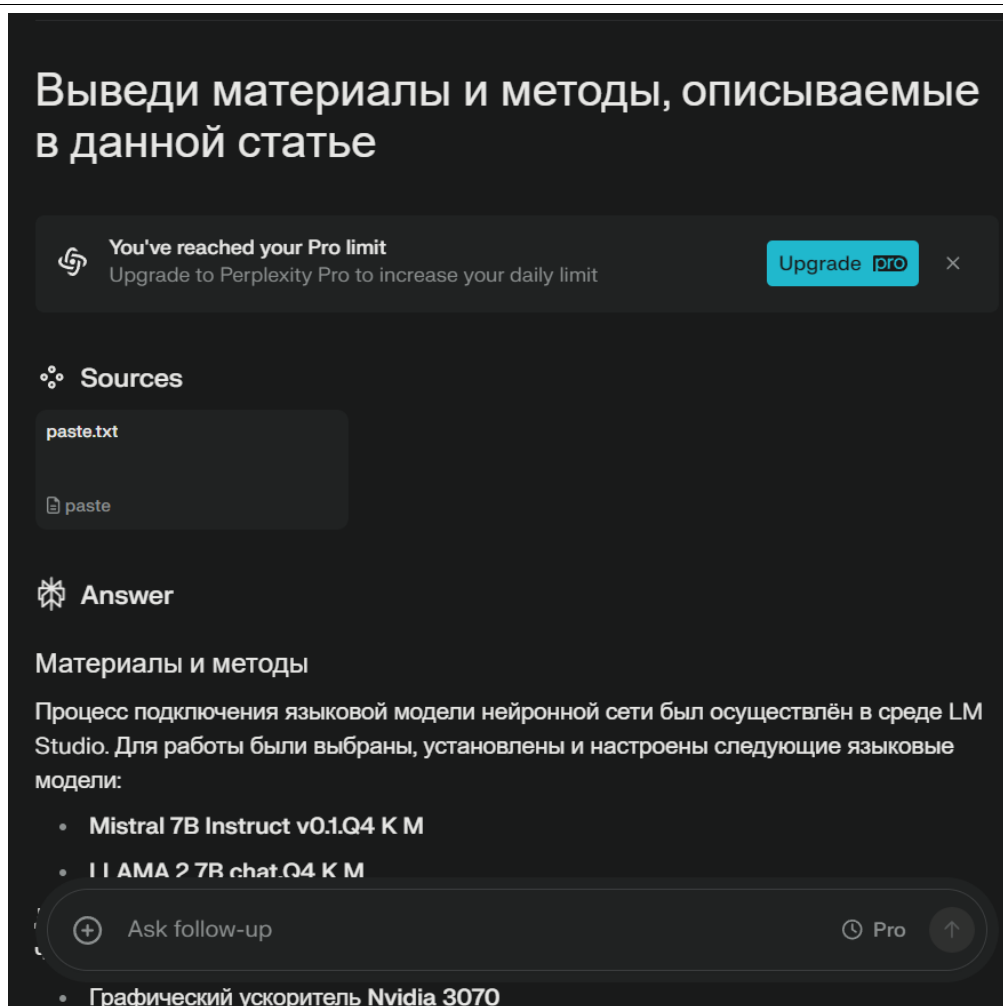


Рисунок 20. Вывод результата запроса информации по тексту

Полученный результат содержал только необходимую информацию напрямую из текста.

4 Выводы

Исследование показало, что работа с языковыми моделями нейронных сетей на локальной машине имеет как преимущества (контроль над данными, безопасность, возможность оффлайн работы), так и проблемы (необходимость мощной вычислительной техники, низкая точность ответа и программные сложности).

LM Studio - потенциально удобный инструмент для работы с языковыми моделями, доступными на Hugging Face. Имеется возможность взаимодействия с моделями через чат-интерфейс. Возможность загрузки и настройки различных параметров работы моделей.

Основным недостатком работы моделей LM Studio является несовершенство инструмента подгрузки файлов, модель некорректно обрабатывает текст из загруженных файлов.

Несмотря на некоторые проблемы с обработкой текста из файлов, модель демонстрирует способность к пониманию и интерпретации информации из текстовых источников.

Исходя из вышеописанного, если нет ограничений в обеспечении излишней конфиденциальности работы, зависимости от наличия интернет соединения, то онлайн сервисы будут более хорошим решением, чем локально развернутая нейросеть, так как нет зависимости от мощности железа и необходимости какой-либо настройке загрузки дополнительных компонентов (моделей).

Библиографический список

1. Павлов Д. А. Искусственные нейросети в контексте науки и образования //Компьютерные инструменты в образовании. 2017. №. 6. С. 25-30.
2. Багаев И. И. Анализ понятий нейронная сеть и сверточная нейронная сеть, обучение сверточной нейросети при помощи модуля TensorFlow // Математическое и программное обеспечение систем в промышленной и социальной сферах. 2020. Т. 8. №. 1. С. 15.
3. Плотникова А. М. Нейросеть как ключевое слово текущего момента //Филологический класс. 2023. Т. 28. №. 2. С. 45-54.
4. Черниговская Т. В. Человеческое в человеке: сознание и нейронная сеть //Проблема сознания в философии и науке. 2008. Т. 20086.
5. LM Studio URL: <https://lmstudio.ai>
6. Perplexity URL: <https://www.perplexity.ai>
7. Hugging Face URL: <https://huggingface.co/>
8. Lmstudio-community/gemma-2-9b-it-GGUF URL: <https://huggingface.co/lmstudio-community/gemma-2-9b-it-GGUF>
9. Звайгзне А. Ю. Развертывание и настройка собственной языковой модели нейронной сети на локальной машине //Постулат. 2024. №. 1. URL: <http://e-postulat.ru/index.php/Postulat/article/view/5429/5515>