МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого»

Институт компьютерных наук и технологий Высшая школа искусственного интеллекта

ОТЧЁТ

Введение в профессиональную деятельность

Студент, Лусиана Андреа Лимпиас Замбрана

Чикаге Майкл

Группы, 5130203/4002

Преподаватель,

Contents

[INSTRUCCIONES 3](#_Toc200396266)

[1) Введение в LLM (языковые модели) 3](#_Toc200396267)

[1.2) Тестирование на данных MMLU 3](#_Toc200396268)

[1.2.1) proceso para poder realizar las preguntas 3](#_Toc200396269)

[1.2.3)respuestas con deepseek 3](#_Toc200396270)

[1.3) Задача 3](#_Toc200396271)

[DESARROLLO 6](#_Toc200396272)

[1.2.4) llama3.2:3b 19](#_Toc200396273)

[**2.1)proceso para poder realizar las preguntas** 20](#_Toc200396274)

# INSTRUCCIONES

# 1) Введение в LLM (языковые модели)

Следуя инструкциям из файла deepseek\_installer.txt (из lab01), выполните на локальном компьютере следующие модели:

DeepSeek-R1:1.5b (версия модели для логических задач с 1,5 миллиардами параметров)

Llama3.2:3b (предобученная модель с квантованием Q4)

Запустите модели в трех сценариях:  
 1.a) локальный запуск в терминале  
 1.b) локальный запуск с использованием Docker-контейнера + Web UI  
 1.c) локальный запуск через LM Studio Desktop

## 1.2) Тестирование на данных MMLU

### 1.2.1) proceso para poder realizar las preguntas

1.2.2)question from dataset

Используя стандартный MMLU Dataset для бенчмаркинга, выберите 3 вопроса:

1 по анатомии

1 по информатике

1 по математике

Ссылка на датасет: <https://www.kaggle.com/datasets/lizhecheng/mmludataset>

### 1.2.3)respuestas con deepseek

1.2.4)respuestas con ollama

## 1.3) Задача

Сравните ответы, полученные от DeepSeek-R1:1.5b и Llama3.2:3b.  
Создайте отчет со скриншотами выполнения запросов в каждом сценарии. Оформите его в виде таблицы с двумя столбцами, где сравниваются ответы каждой модели на каждый вопрос, и прокомментируйте, какая модель дает более точные результаты.

2) Инженерия подсказок с использованием RAG

Возьмите проект **Web-Chatbot** (опубликован на GitHub — lab02), разработанный с использованием **LLM и RAG**, и с учетом следующей структуры папок и файлов выполните запуск приложения локально на компьютере:

Ссылка на репозиторий:  
 <https://github.com/HoltechHard/Professional_Proj_Activity-40001-2>

**Структура проекта:**

* **faiss\_db:** папка, содержащая базу данных векторных эмбеддингов.
* **history:** папка, содержащая историю взаимодействий «вопрос-ответ» в чат-боте.
* **scrap:** папка, содержащая текстовую информацию, полученную при скрейпинге веб-сайта.
* **scrapper.py**: скрипт Python для сбора и скрейпинга контента с веб-сайта.
* **summarization.py:** скрипт Python для генерации резюме на основе скрейпинга.
* **ingest.py**: скрипт Python для построения векторной базы данных из кусков скрейп-данных.
* **chatbot.py:** скрипт Python, содержащий процессы индексации, извлечения и генерации ответа на заданный вопрос.
* **app.py:** основной скрипт Python, реализующий интеграцию Streamlit с бэкендом Langchain и RAG.  
  Для запуска приложения выполните команду:
* **streamlit run app.py**
* **requirements.txt**: файл, содержащий список необходимых Python-библиотек.  
  Для установки выполните команду:
* **pip install -r requirements.txt**

Задачи:

* 1. **Адаптировать исходный код** приложения Web-Chatbot так, чтобы пользователь мог выбрать **одну из 4 LLM моделей** от разных провайдеров (например: deepseek, qwen, gpt-4o, llama).
  2. Процессы **скрейпинга и создания эмбеддингов не изменяются**.  
     В модулях **summarization.py** и **chatbot.py** необходимо добавить **комбо-бокс** с выбором модели, и в зависимости от выбора пользователь должен получать резюме или ответ, сформированные соответствующей LLM.
  3. Выберите **один технический веб-сайт**, содержащий достаточный объем специализированной информации.  
     Проведите:
* 1 суммаризацию
* 1 вопрос-ответ  
  для каждой из 4 выбранных моделей.

Реализуйте функцию Python, которая **измеряет время выполнения (в секундах)** от начала до получения ответа.  
Постройте график **benchmark** по примеру из изображения (ось Y — время, ось X — модели).

* 1. Составьте **отчет** по **качеству ответов каждой LLM** на один технический вопрос.  
     Включите **таблицу сравнений** с анализом ответов каждой модели. Укажите:
* Точность
* полноту
* релевантность
* стиль объяснения и пр.

# DESARROLLO

1) Введение в LLM (языковые модели)

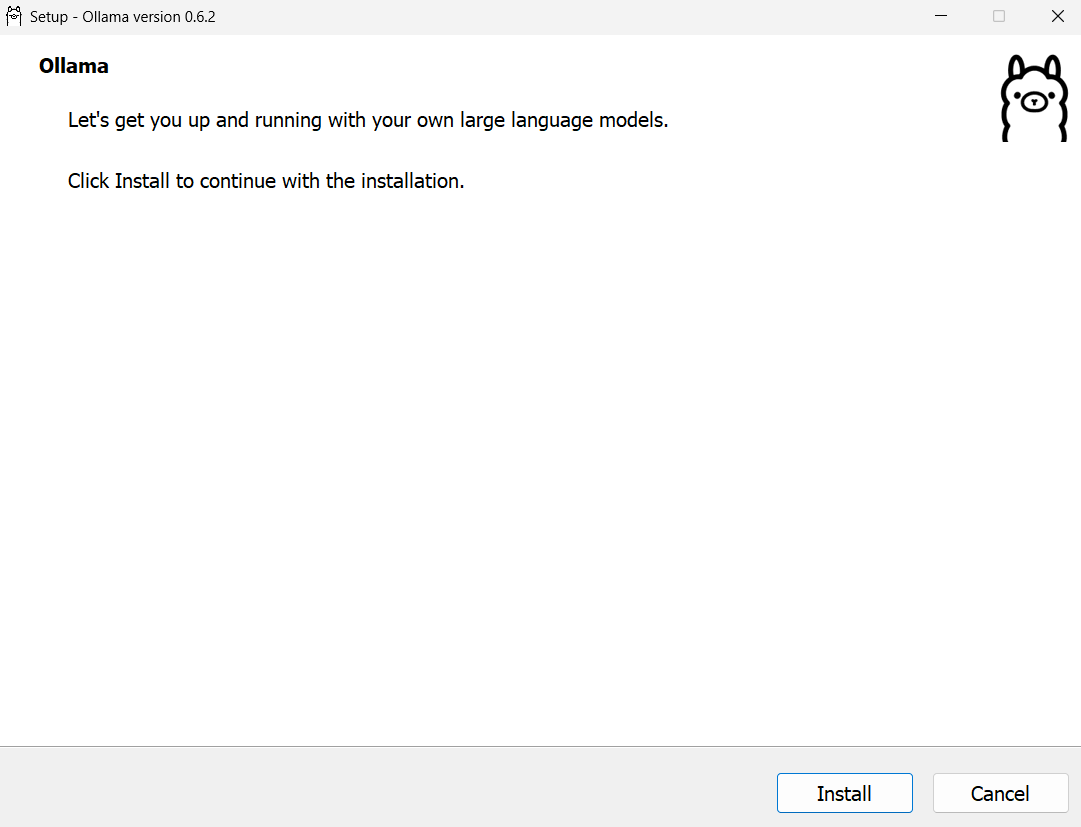
1.a) локальный запуск в терминале

-Перейти на официальный сайт ollama: <https://ollama.com>

A screenshot of a computer

AI-generated content may be incorrect.

- Запустите установщик для Windows и выполните установку



- Проверьте, что Ollama успешно установился

A computer screen shot of a program

AI-generated content may be incorrect.

- Скачайте запрошенные модели: llama3-2:3b и deepseek-r1:1.5b A screenshot of a computer

AI-generated content may be incorrect.

Запустите в Windows PowerShell следующую команду:

-Запустите модели из консоли командой:  
ollama run <модель>

A black screen with white text

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.

1.b) локальный запуск с использованием Docker-контейнера + Web UI

Перейти на официальный сайт Docker: <https://docker.com>

A screenshot of a computer

AI-generated content may be incorrect.

Запуск контейнера Open WebUI с помощью команд терминала

official web site documentation of open webui: <https://docs.openwebui.com/getting-started/quick-start/>

A screenshot of a computer

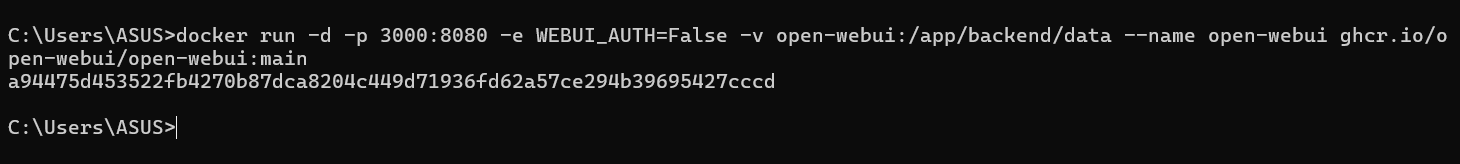
AI-generated content may be incorrect.

- Скачайте образ Open WebUI

A black background with white text

AI-generated content may be incorrect.

- Запустите контейнер Open WebUI с параметром WEBUI\_AUTH=False (без требования авторизации)



- Проверьте, что контейнер запущен на порту 3000

A screenshot of a computer

AI-generated content may be incorrect.

**API Ollama: отображение моделей**

<https://github.com/ollama/ollama/blob/main/docs/api.md>

A computer screen with many small colorful text

AI-generated content may be incorrect.

Откройте браузер и перейдите на порт 3000, чтобы проверить загруженные модели Ollama

A screenshot of a computer

AI-generated content may be incorrect.

Взаимодействие с моделями через веб-интерфейс

A screenshot of a computer

AI-generated content may be incorrect.

1.c) локальный запуск через LM Studio Desktop

- Скачайте LM Studio Desktop с официальной страницы: <https://lmstudio.ai>

A screenshot of a computer

AI-generated content may be incorrect.

- Загрузка моделей для использования в LM Studio

A screenshot of a computer

AI-generated content may be incorrect.

Выбор модели для использования в LM Studio

A screenshot of a computer

AI-generated content may be incorrect.

Взаимодействие с моделью через графический интерфейс LM Studio

A screenshot of a computer

AI-generated content may be incorrect.

1.2) Тестирование на данных MMLU

1.2.3)preguntas de dataset

**1 по анатомии**

A screenshot of a computer

AI-generated content may be incorrect.

The infraorbital nerve:

1. is a terminal branch of the maxillary division of the trigeminal nerve.
2. is a terminal branch of the maxillary branch of the facial nerve.
3. carries parasympathetic secretomotor neurons to the lacrimal gland.
4. innervates the inferior part of the orbicularis oculi muscle.

The correct answer is **a**

**1 по информатике**

A screenshot of a computer

AI-generated content may be incorrect.

Which of the following is true of interrupts?

1. They are generated when memory cycles are "stolen".
2. They are used in place of data channels.
3. They can indicate completion of an I/O operation.
4. They cannot be generated by arithmetic operations.

Correct answers **c**

**1 по математике**

A screenshot of a calendar

AI-generated content may be incorrect.

The map x -> axa^2 of a group G into itself is a homomorphism if and only if

1. G is abelian

b) G = {e}

c) a^3 = e

d) a^2 = a

correct answer **c**

1.2.3)DeepSeek-R1.5b

**1 по анатомии**

**A screenshot of a computer

AI-generated content may be incorrect.**

**1 по информатике**

A black screen with white text

AI-generated content may be incorrect.

**1 по математике**

A black screen with white text

AI-generated content may be incorrect.

**1.2.4) llama3.2:3b**

**1 по анатомии**

A black screen with white text

AI-generated content may be incorrect.

**1 по информатике**

A screen shot of a computer

AI-generated content may be incorrect.

**1 по математике**

A screenshot of a computer

AI-generated content may be incorrect.

### 2.1)proceso para poder realizar las preguntas

Скачать пример проекта из репозитория:

<https://github.com/HoltechHard/Professional_Proj_Activity-40001-2>

Для работы примера проекта необходим Python и библиотеки из файла requirements.txt

### 

A black screen with white text

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

Я использую Python 3.13.2

C:\Users\ASUS\AppData\Local\Programs\Python\Python313\Scripts

Запустить программу (веб-версия)

A black screen with a black background

AI-generated content may be incorrect.

C:\Users\ASUS\AppData\Local\Programs\Python\Python313\Scripts\streamlit run app.py

A screenshot of a computer

AI-generated content may be incorrect.

Для получения списка моделей использовался API Ollama по адресу: http://{host:port}/api/tags

A computer screen shot of text

AI-generated content may be incorrect.

Для работы библиотеки веб-скрапинга требуется установка Chromium

### 

A black background with white text

AI-generated content may be incorrect.

2.2)

Веб-сайт был адаптирован, чтобы пользователь мог выбирать между моделями, загруженными в сервис Ollama.

A screenshot of a computer

AI-generated content may be incorrect.

Класс WebSummarizer был изменён - теперь модель указывается в конструкторе

No te olvides de explicar luego el código y porque lo hiciste

A screen shot of a computer

AI-generated content may be incorrect.

Функция summarize была изменена - теперь она показывает время ответа модели в секундах.

A screen shot of a computer program

AI-generated content may be incorrect.

В классе Chatbot был добавлен параметр model в конструктор - это модель Ollama для использования.

A screen shot of a computer program

AI-generated content may be incorrect.

Функция qa была изменена - теперь она возвращает время ответа.

A computer screen shot of text

AI-generated content may be incorrect.

2.3) Executing the summarization

For this purpose, we selected website <https://english.spbstu.ru/university>

and it was executed with every model, the models that we will use are: deepseek-r1:1.5b, qwen2.5:latest, llama3.2:3b, deepseek-r1:8b.

The question will be “Which is the address of the university?”

Selected side:

A screenshot of a computer

AI-generated content may be incorrect.

2.3.1) Working with **qwen2.5**

**a) Web scrapping**

A screenshot of a chatbot

AI-generated content may be incorrect.

Realiza el proceso de web scrapping y web summarization, en el modelo **qwen2.5** tarda **21.12** segundos

**b) Web Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

La pregunta-respuesta, en el modelo **qwen2.5** tarda **1.98** segundos y calidad de **3**

**c) ChatBot**

A screenshot of a chatbot

AI-generated content may be incorrect.

2.3.2) Working with llama3.2:3b

**a) Web scrapping**

A screenshot of a computer

AI-generated content may be incorrect.

**b) Web Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

La pregunta-respuesta, en el modelo **llama3.2:3b** tarda **2.29** segundos y calidad de **5**

**c) ChatBot**

A screenshot of a computer

AI-generated content may be incorrect.

2.3.3) Working with deepseek-r1:1.5b

**a) Web scrapping**

A screenshot of a computer

AI-generated content may be incorrect.

**b) Web Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

La pregunta-respuesta, en el modelo **deepseek-r1:1.5b** tarda **3.18** segundos y calidad de **3.8**

**c) ChatBot**

A screenshot of a computer

AI-generated content may be incorrect.

2.3.5) Working with llama3.2:3b

**a) Web scrapping**

A screenshot of a computer

AI-generated content may be incorrect.

**b) Web Summarization**

A screenshot of a computer

AI-generated content may be incorrect.

La pregunta-respuesta, en el modelo **deepseek-r1:8b** tarda **9.10** segundos y calidad de **4**

**c) ChatBot**

A screenshot of a computer

AI-generated content may be incorrect.

2.4)final results

A close-up of a computer screen

AI-generated content may be incorrect.

A graph with blue bars

AI-generated content may be incorrect.

A graph with blue squares

AI-generated content may be incorrect.