Министерство науки и высшего образования Российской Федерации федеральное государственное автономное образовательное учреждение высшего образования

«Санкт-Петербургский политехнический университет Петра Великого» (ФГАОУ ВО СПбПУ)

Институт «Компьютерных наук и кибербезопасности»

**Введение в профессиональную деятельность**

Введение в LLM и Prompt Engineering с RAG

(Семестр 1)

**Студенты группы**: 5130203/40002 Лусиана Андреа Лимпиас Замбрана

Майкл Чиканге

**Преподаватель:** Хольгер Эспинола Ривера

Санкт-Петербург – 2025

»

**1.Введение в LLM**

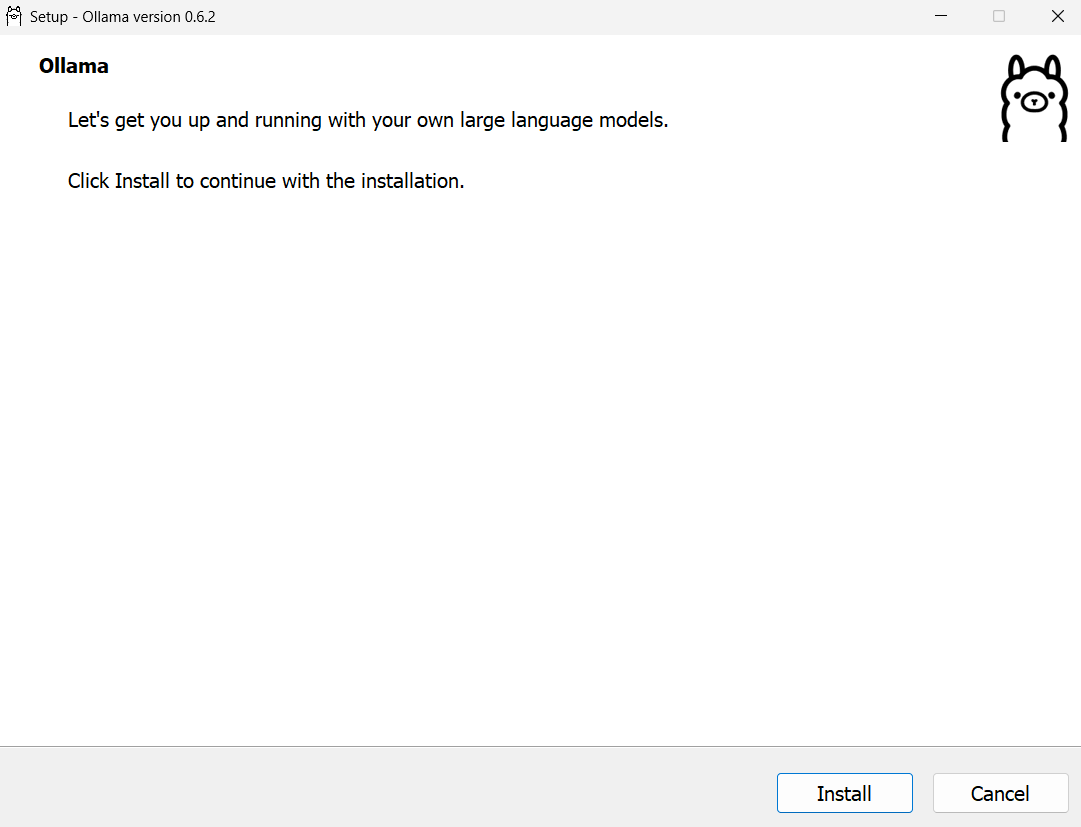
1. Установить Ollama

- Перейдите 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

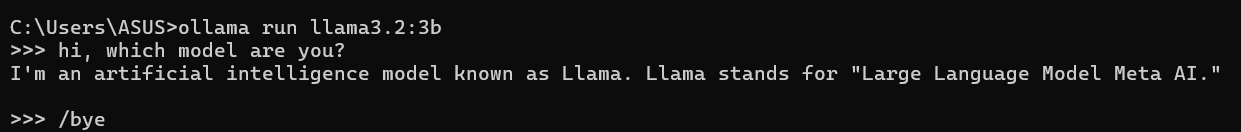
A screenshot of a computer

AI-generated content may be incorrect.

**1.a) запуск локально в терминале**Запустите модели из консоли командой:  
ollama run <модель>

A black screen with white text

AI-generated content may be incorrect.



**1.b) запуск локально с использованием контейнера Docker + веб-интерфейса**Установите и запустите Docker с официального сайта: https://docs.docker.com/desktop/setup/install/windows-install/

A screenshot of a computer

AI-generated content may be incorrect.

Следуйте шагам из документации официального сайта Open WebUI

https://docs.openwebui.com/getting-started/quick-start/

A screenshot of a computer

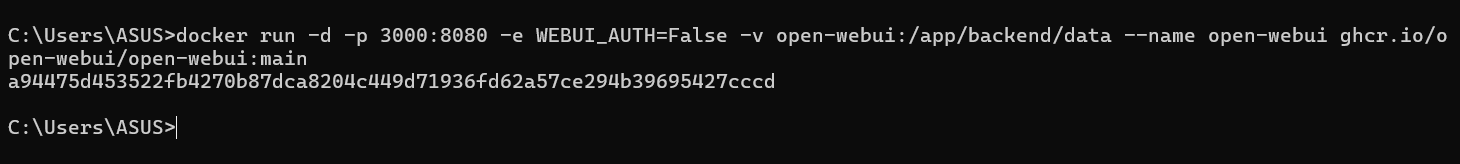
AI-generated content may be incorrect.

- 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)** **запуск локально с использованием LM Studio Desktop**

Установка LM Studio

<https://lmstudio.ai/download>

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.

**2.Оперативное проектирование с RAG.**

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 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.

**Задача:**

1)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

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) Процессы суммаризации и работы чатбота выполняются в соответствии с выбранной моделью.**

**3)**

**4)**

**Для тестирования был выбран сайт:**

<https://english.spbstu.ru/university>

A screenshot of a chatbot

AI-generated content may be incorrect.

Процесс веб-скрапинга и суммаризации выполнен - модель Qwen2.5 обработала запрос за 21.12 секунды.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Вопрос-ответ: модель Qwen2.5 обрабатывает запрос за 1.98 секунды с качеством 3.

A screenshot of a chatbot

AI-generated content may be incorrect.

Процесс веб-скрапинга и суммаризации выполнен — модель Llama3.2:3b обработала запрос за 11.51 секунды

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Модель Llama3.2:3b обрабатывает вопрос-ответ за 2.29 секунды с максимальным качеством (5/5).

A screenshot of a computer

AI-generated content may be incorrect.

Модель DeepSeek-R1:1.5b выполняет веб-скрапинг и суммаризацию за 11.47 секунд.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Модель DeepSeek-R1:1.5b обрабатывает вопрос-ответ за 3.18 секунды с качеством 3.8/5.

A screenshot of a computer

AI-generated content may be incorrect.

Модель DeepSeek-R1:8b выполняет веб-скрапинг и суммаризацию за 35.23 секунды.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Модель DeepSeek-R1:8b обрабатывает вопрос-ответ за 9.10 секунд с качеством 4/5.

A screenshot of a computer

AI-generated content may be incorrect.

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.

A graph with blue bars

AI-generated content may be incorrect.