Contents

[1. Introduction to LLM 2](#_Toc200100572)

[1.1) Installation 3](#_Toc200100573)

[1.2) Questions from Dataset 8](#_Toc200100574)

[1.3) Answering with DeepSeek-R1:1.5b 11](#_Toc200100575)

[1.4) Answering with Llama3.2:3b 13](#_Toc200100576)

[1.5) Comparing results 14](#_Toc200100577)

[2) Prompt Engineering with RAG. 15](#_Toc200100578)

[2.1) Installation 17](#_Toc200100579)

[2.2) Modification of code to choose models 18](#_Toc200100580)

[2.3) Executing the summarization 19](#_Toc200100581)

[2.3.1) Working with deepseek-r1:1.5b 19](#_Toc200100582)

[2.3.2) Working with qwen2.5:latest 22](#_Toc200100583)

[2.3.3) Working with llama3.2:3b 25](#_Toc200100584)

[2.3.4) Working with deepseek-r1:8b 27](#_Toc200100585)

# 1. Introduction to LLM

Following the instructions contained in document deepseek\_installer.txt (from lab01), execute in local computer the next models:

\* DeepSeek-R1:1.5b (version of reasoning model with 1.5 billion of parameters)

\* Llama3.2:3b (version of pre-trained model with quantization Q4)

Execute the models in the next 3 scenarios:

a) running locally in terminal

b) running locally using Docker container + Web UI

c) running locally using LM Studio Desktop

Taking in account the standard MMLU Dataset for benchmarks, select 3 questions: 1 for anatomy test, 1 for computer science and 1 for mathematics. Use the link bellow as reference: <https://www.kaggle.com/datasets/lizhecheng/mmludataset>

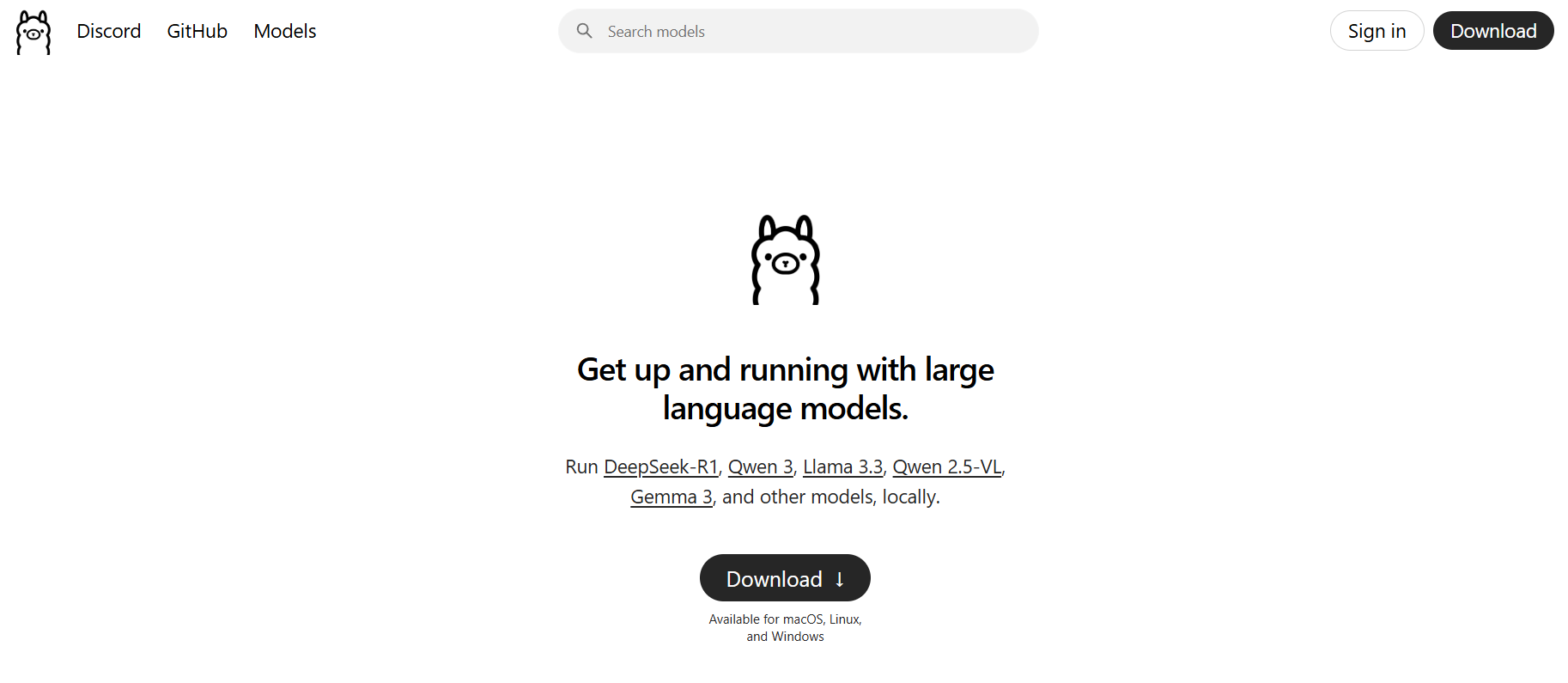
**Task:** Compare the answers obtained by DeepSeek-R1:1.5b and Llama3.2:3.2.

Make a report about it with screen shots of execution of prompts in each scenario. You can do it as a two-column table comparing each answer obtained for each question comparing the 2 models and comment what model performs better answers.

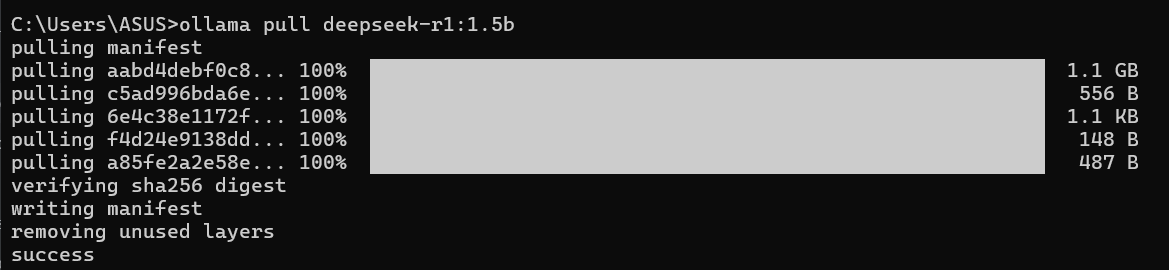
## 1.1) Installation

**a) running locally in terminal**

Go to the official web site of ollama: <https://ollama.com>



Download DeepSeek-R1-8b model. It requires 4.9 GB of RAM memory



Run in Windows PowerShell the next command:

$ ollama run deepseek-r1:1.8b

**b) running locally using Docker container + Web UI**

Go to the official site of docker: <https://docker.com>

A screenshot of a computer

AI-generated content may be incorrect.

Run Open WebUI container: using terminal commands

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



Run the Image from docker

docker run -d -p 3000:8080 -e WEBUI\_AUTH=False -v open-webui:/app/backend/data --name open-webui ghcr.io/open-webui/open-webui:main

A screenshot of a computer

AI-generated content may be incorrect.

Try the model

A screenshot of a computer

AI-generated content may be incorrect.

**c) running locally using LM Studio Desktop**

Download LM Studio Desktop from the official page: <https://lmstudio.ai>

A screenshot of a computer

AI-generated content may be incorrect.

Download the model from the browser

A screenshot of a computer

AI-generated content may be incorrect.

Try the model

A screenshot of a computer

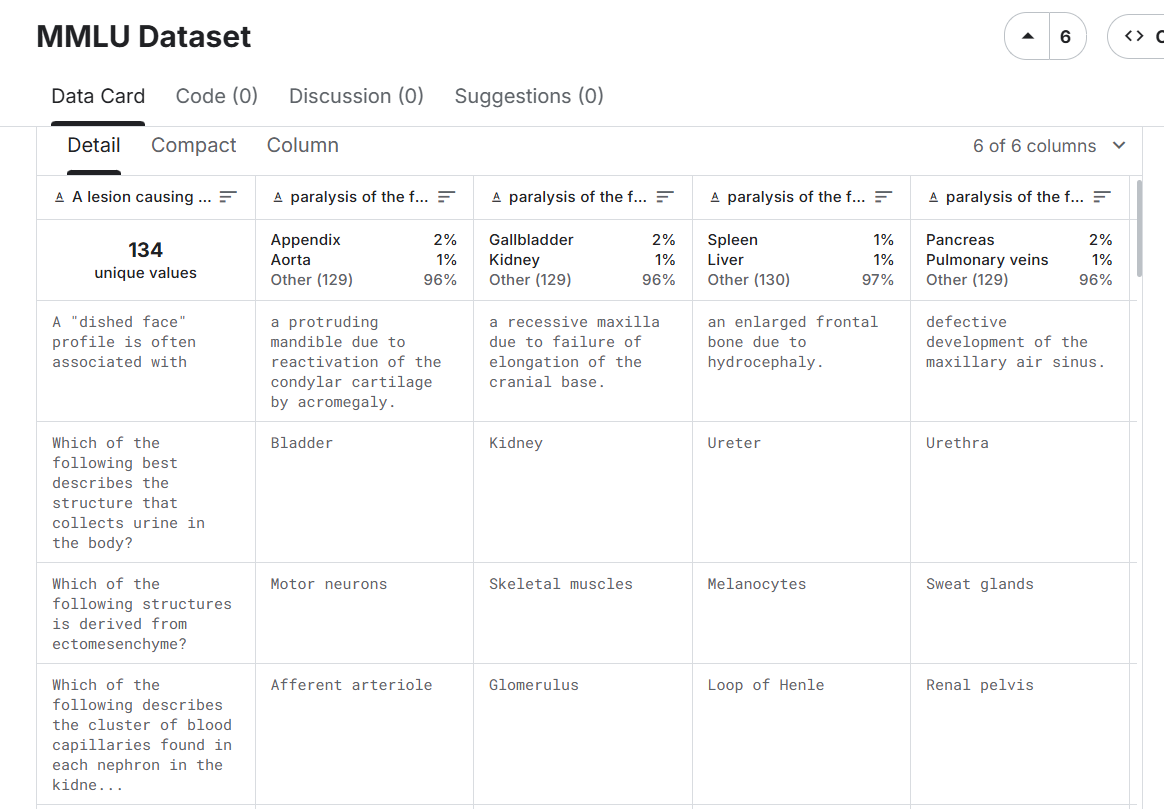
AI-generated content may be incorrect.

## 1.2) Questions from Dataset

For testing the different models we will use the next dataset of questions for anatomy, computer science and mathematics.

**DATASETS:**

**Anatomy**



1. Which of the following anatomical regions of abdomen lies just distal to the sternum?

A) Epigastric

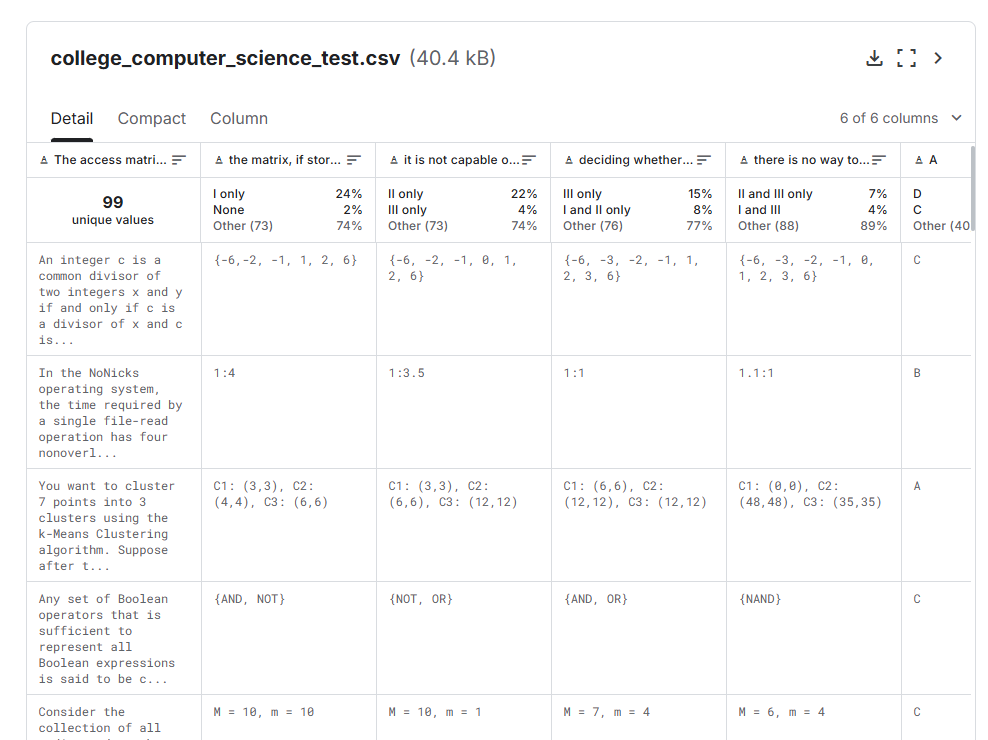
B) Hypochondriac

C) Hypogastric

D) Lumbar

Correct Answer: **A**

**Computer Science**



2. The language {ww | w in (0 + 1)\*} is

A) not accepted by any Turing machine

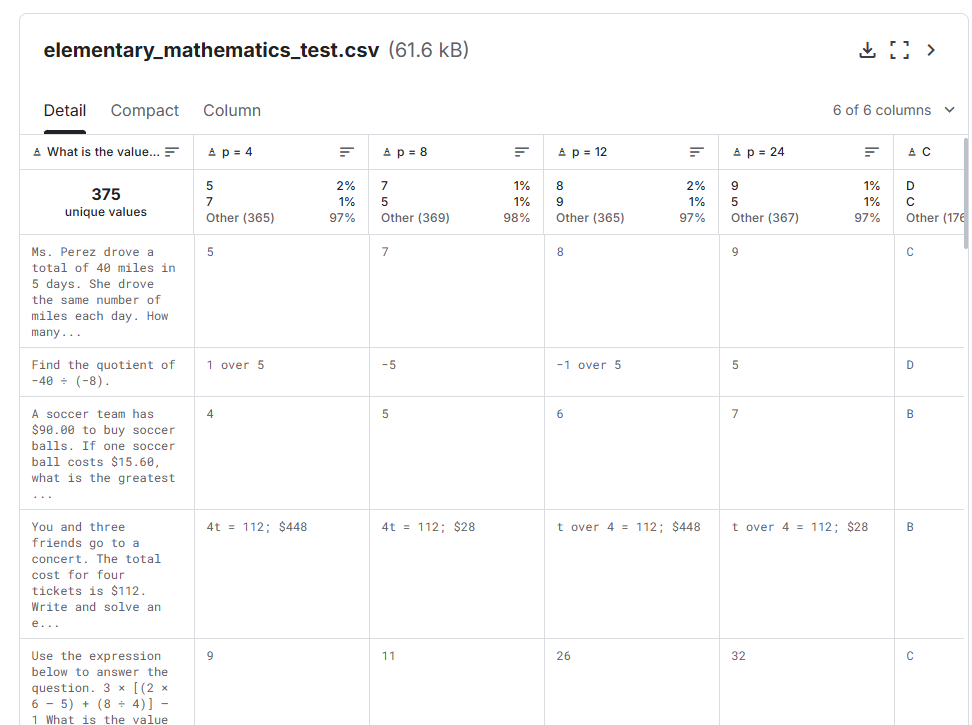
B) accepted by some Turing machine, but by no pushdown automaton

C) accepted by some pushdown automaton, but not context-free

D) context-free, but not regular

Correct Answer: **B**

**Mathematics**



3. Write 1 and 1 over 11 as a whole or mixed number in simplest form.

A) 1.1 over 11

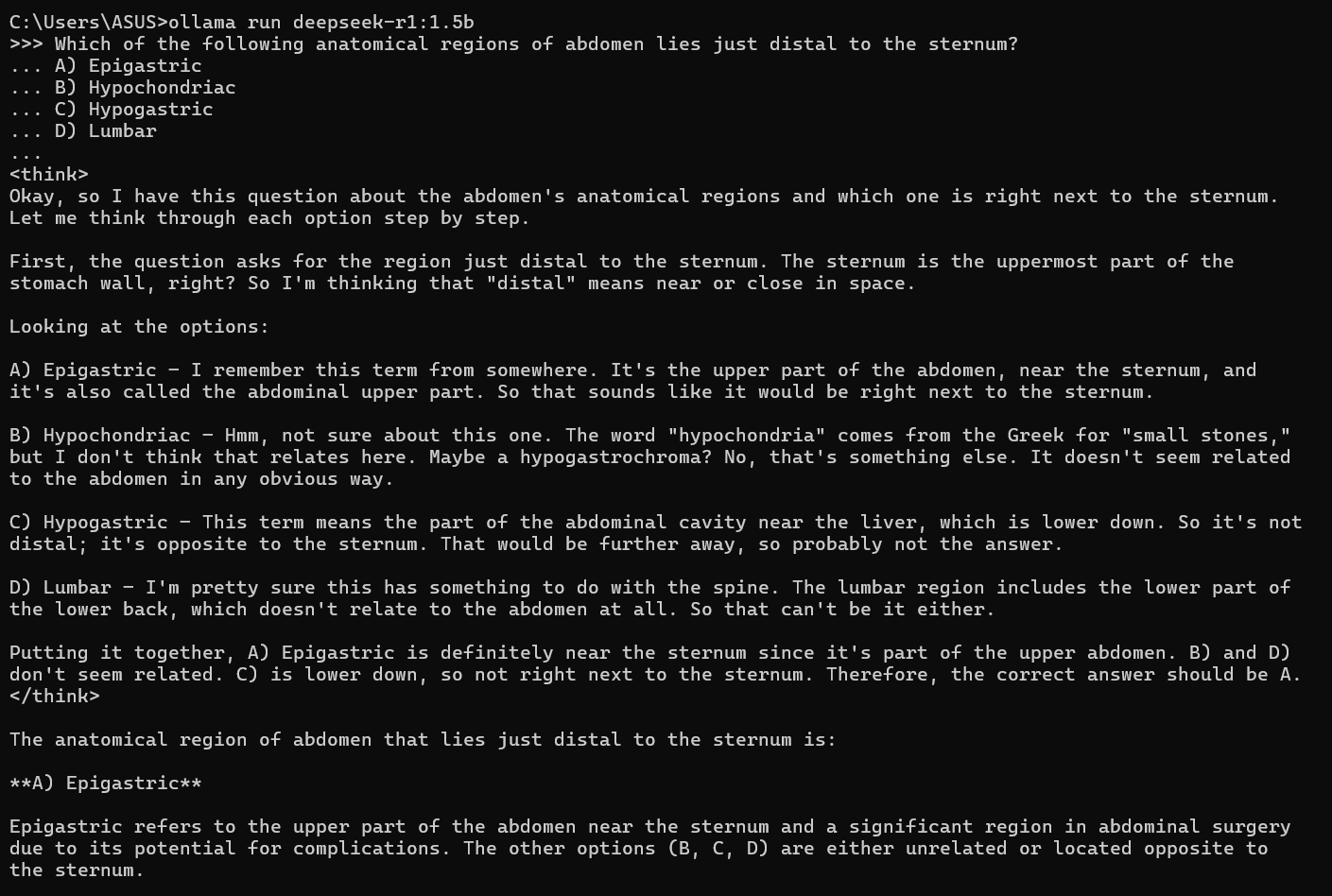
B) 11 over 11

C) 12 over 11

D) 11 over 12

Correct Answer: **C**

## 1.3) Answering with DeepSeek-R1:1.5b

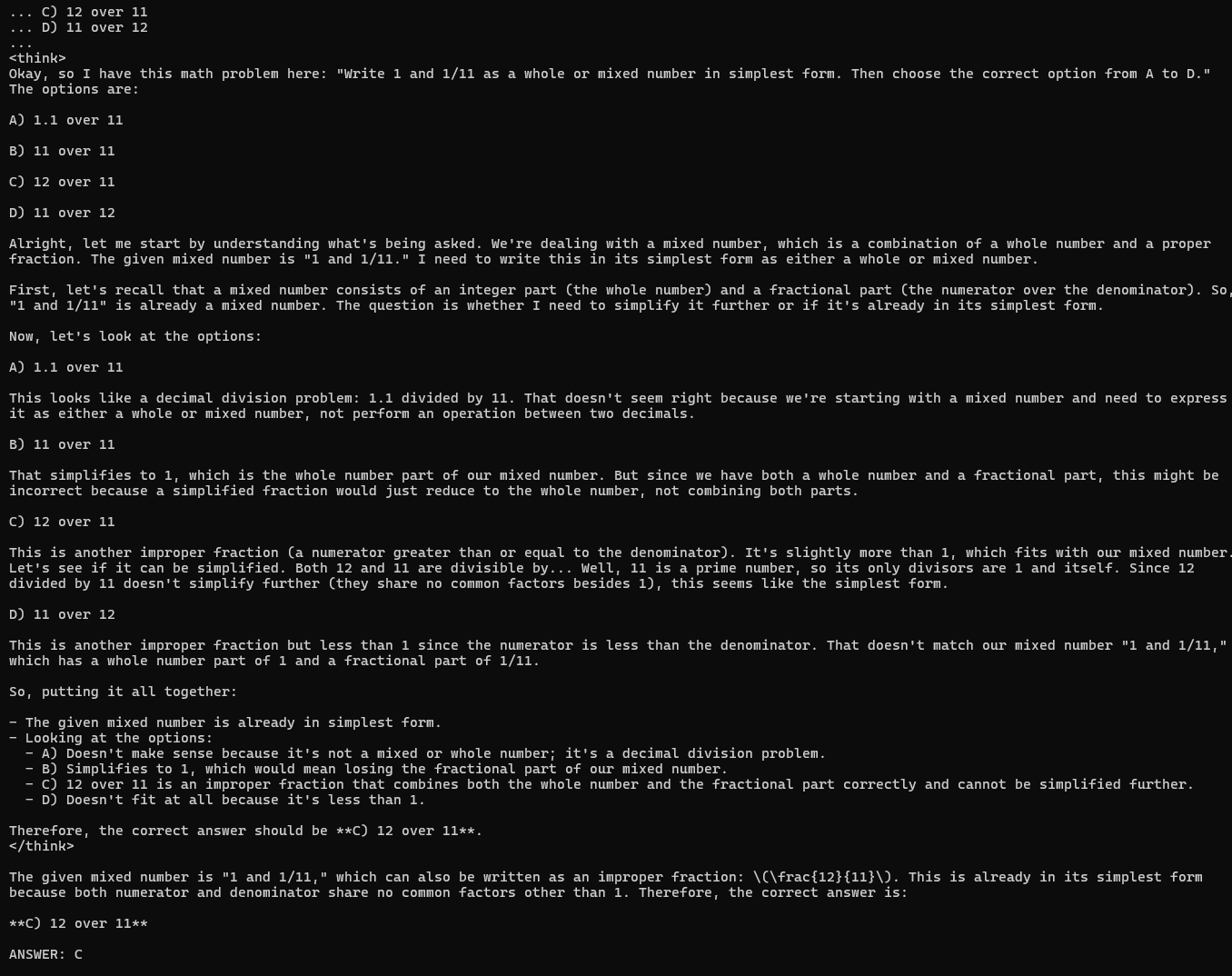


The IA responses the option A (Correct)

A black screen with white text

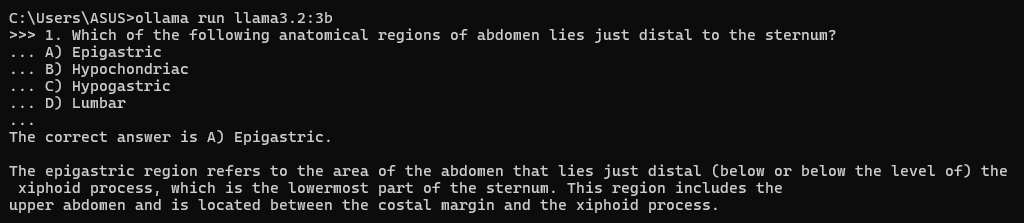
AI-generated content may be incorrect.

The IA responses the option D (Incorrect)



The IA responses the option C (Correct)

## 1.4) Answering with Llama3.2:3b



The IA responses the option A (Correct)

A screenshot of a computer program

AI-generated content may be incorrect.

The IA responses the option D (Incorrect)

A screenshot of a computer program

AI-generated content may be incorrect.

The IA responses the option None? (Incorrect)

## 1.5) Comparing results

|  |  |  |
| --- | --- | --- |
| **Model** | **DeepSeek-R1:1.5b** | **Llama3.2:3b** |
| Correct Answers? (1-3) | 2 | 1 |
| Quality of results (1-5) | 5 | 3 |
| Length of results (1-5) | 5 | 2 |

The better model from the experiment was DeepSeek-R1:15.b with more correct answers, better qualify of results and better length of results.

# 2) Prompt Engineering with RAG.

Take the project of Web-Chatbot (published in Github - lab02) developed with LLM and RAG and taking in account the next structure of folders and files, execute the application locally in computer: Reference: <https://github.com/HoltechHard/Professional_Proj_Activity-40001-2>

- faiss\_db: folder which contains the database of vector embeddings

- history: folder which contains the history of interactions question-answering in chatbot

- scrap: folder which contains the plain text of scraped information from web site

- scrapper.py: python script to crawl and scrap content from web site

- summarization.py: python script to generate summary of web scraped content

- ingest.py: python script to build a vector database with vector embeddings of chucks of scraped content from web site

- chatbot.py: python script which contain processes of index, retrieve and generate answer from given question to chatbot

- app.py: main python script which contain the integration of streamlit with langchain backend implementation of RAG. To execute application is necessary run the next command:

**$ streamlit run app.py**

- requirements.txt: text file which contains the list of python packages necessary to install in order to run the application appropriately. To install correctly is necessary run the next command:

**$ pip install -r requirements.txt**

**Task:**

1) Adapt the source code of Web-Chatbot application to user of system have possibility to choose 4 LLMs of different providers (for example: deepseek, qwen, gpt-4o, llama, for example).

2) The processes of scraping and embeddings not change. In processes of summarization and chatbot, user need see a combo box with the 4 LLM options and select it, and python scripts need execute the summarization or chatbot according the selected model.

3) Select one web site which contains reasonable volume of technical content. Formulate the summarization and 1 question-answering for each of the 4 LLM selected. Implement a python function to calculate how much time (in seconds) the system spends until obtaining the answer. Draw a benchmark according the figure bellow:

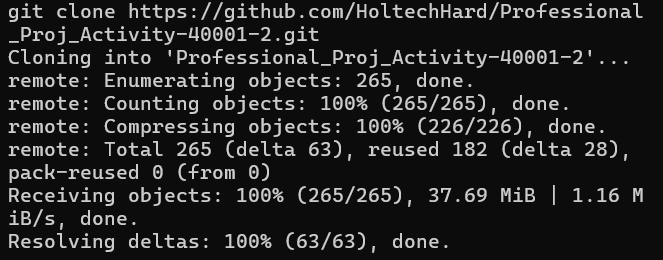
A graph and a diagram

AI-generated content may be incorrect.

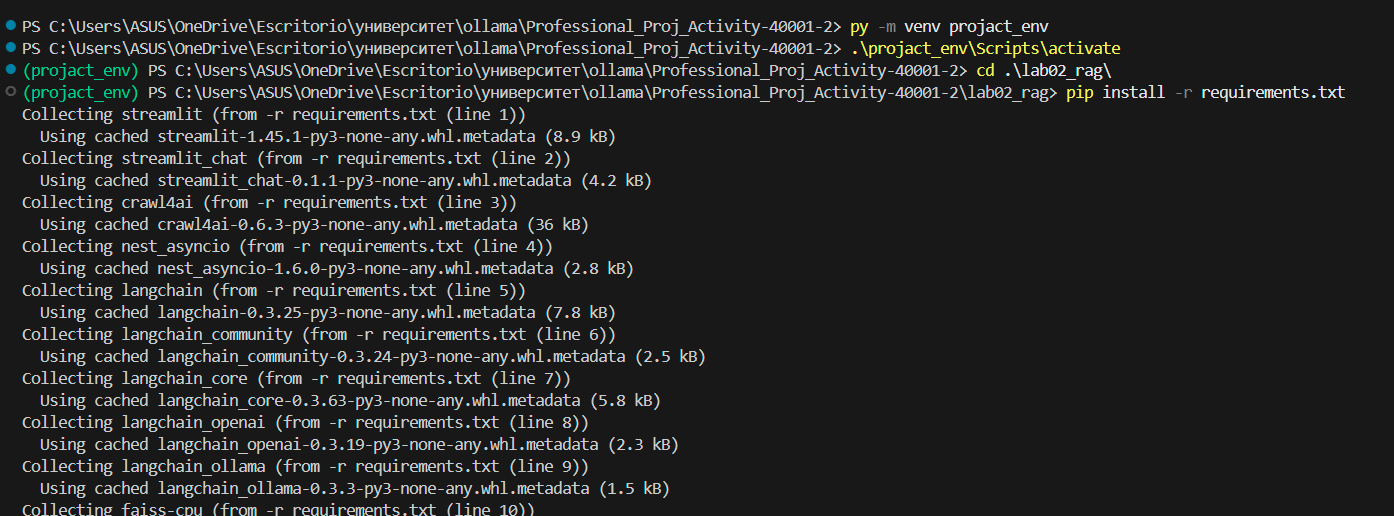
4) Make a report comparing the quality of answers for each LLM answer given a question, according your criteria. Make a table with cross-comparison of the answers of each LLM for 1 technical question related of the web-site scraped.

## 2.1) **Installation**

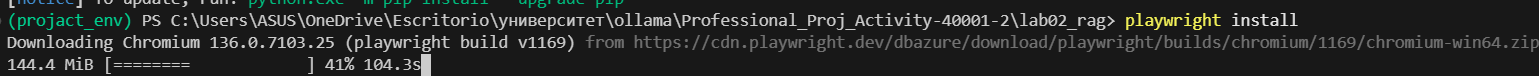
a) Download the project from github



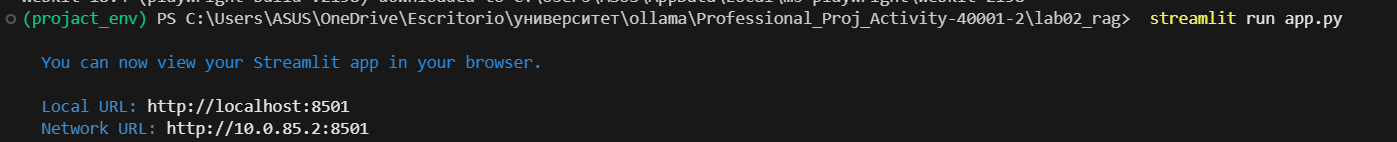
b) Install the dependencies using python environment

****

c) Install playwright

****

d) Run the app

****

**A screenshot of a computer

AI-generated content may be incorrect.**

## 2.2) Modification of code to choose models

For get the models of ollama we will use the API, that list all the current models that have been pulled

Endpoint:

<http://localhost:11434/api/tags>

Were added the models on the classes

A screen shot of a computer code

AI-generated content may be incorrect.

A computer screen with text and numbers

AI-generated content may be incorrect.

To list the models, we added the Models class, that consumes from the API

A screen shot of a computer program

AI-generated content may be incorrect.

Finally we consume the service, and add the model to classes

A black screen with white text

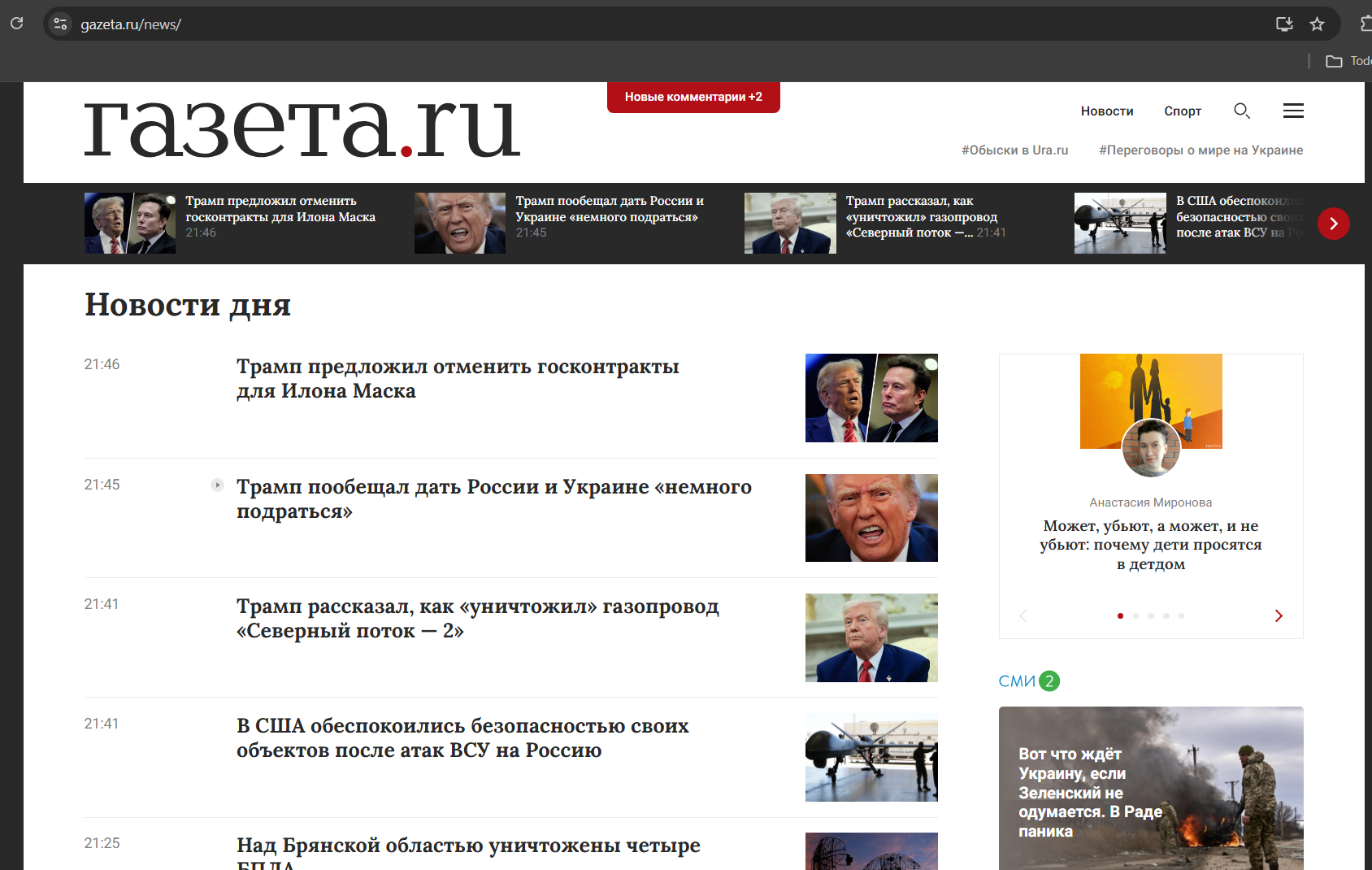
AI-generated content may be incorrect.

## 2.3) Executing the summarization

For this purpose, we selected website <https://www.gazeta.ru/news/> 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 are the last 4 news?”

Selected site



### 2.3.1) Working with deepseek-r1:1.5b

**a) Web scrapping**

A screenshot of a chatbot

AI-generated content may be incorrect.

**b) Web Summarization**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A black screen with white text

AI-generated content may be incorrect.**

**c) ChatBot**

**A screenshot of a computer

AI-generated content may be incorrect.**

### 2.3.2) Working with qwen2.5:latest

**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.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**c) ChatBot**

A screenshot of a computer

AI-generated content may be incorrect.

### 2.3.3) Working with llama3.2:3b

**a) Web scrapping**

A screenshot of a chatbot

AI-generated content may be incorrect.

**b) Web Summarization**

**A screenshot of a computer

AI-generated content may be incorrect.**

**c) ChatBot**

A screenshot of a computer

AI-generated content may be incorrect.

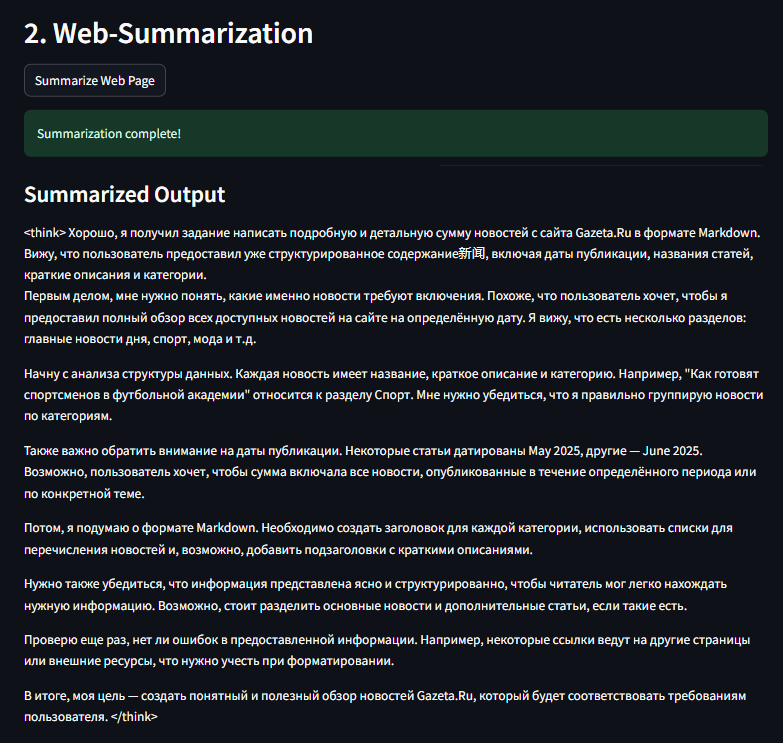
### 2.3.4) Working with deepseek-r1:8b

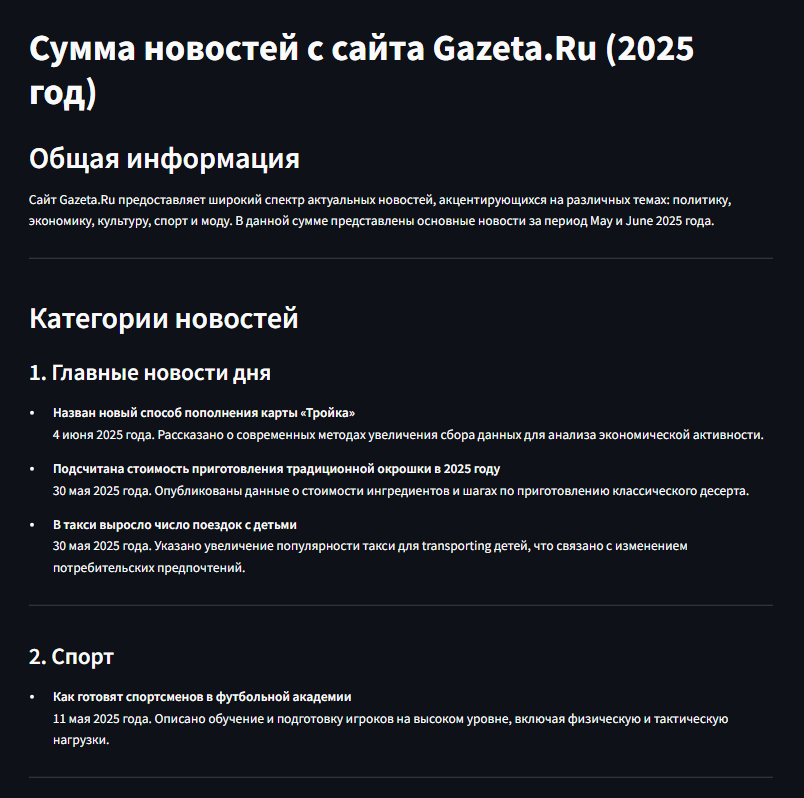
**a) Web scrapping**

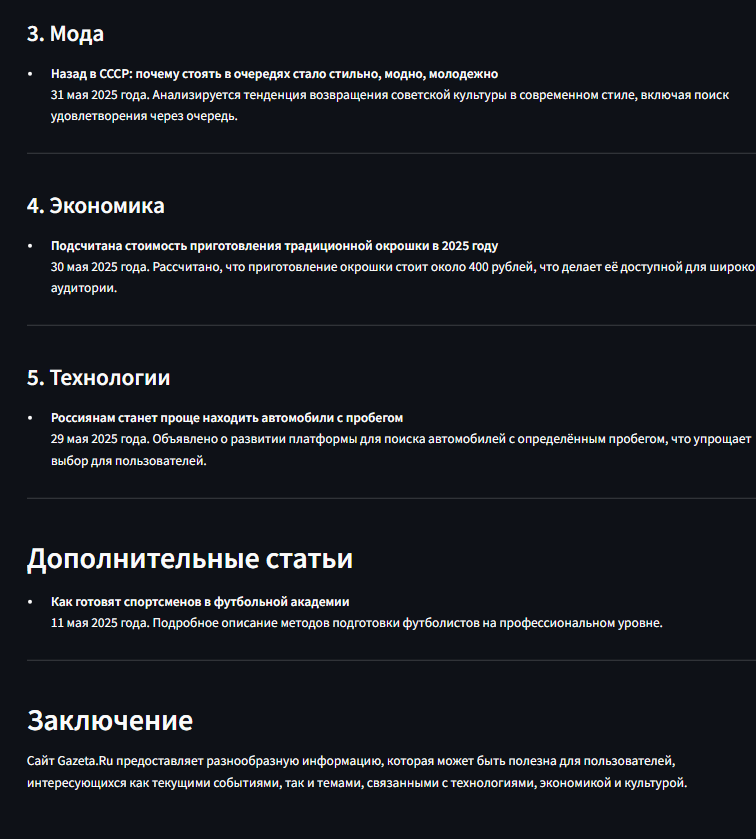
A screenshot of a computer

AI-generated content may be incorrect.

**b) Web Summarization**

****

****

****

**c) ChatBot**

A screenshot of a computer

AI-generated content may be incorrect.

### 2.4) Comparison of Models

We used our criteria to form comparison of models.

|  |  |  |
| --- | --- | --- |
| **Model** | **Time** | **Quality (1-10)** |
| deepseek-r1:1.5b | 4.7 | 10 |
| qwen2.5:latest | 26.76 | 4 |
| llama3.2:3b | 14.1 | 7 |
| deepseek-r1:8b | 130.37 | 8 |

