ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

***«*САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»**

**Институт компьютерных наук и кибербезопасности**

**Домашнее задание №1 по предмету ВвПД Введение в LLM и Prompt Engineering c RAG**

Авторы работы: Студенты группы 5130203/40002

Сиволобов Александр Чумак Артемий

Севоян Ваграм

Голдобин Кирилл

САНКТ-ПЕТЕРБУРГ 2025

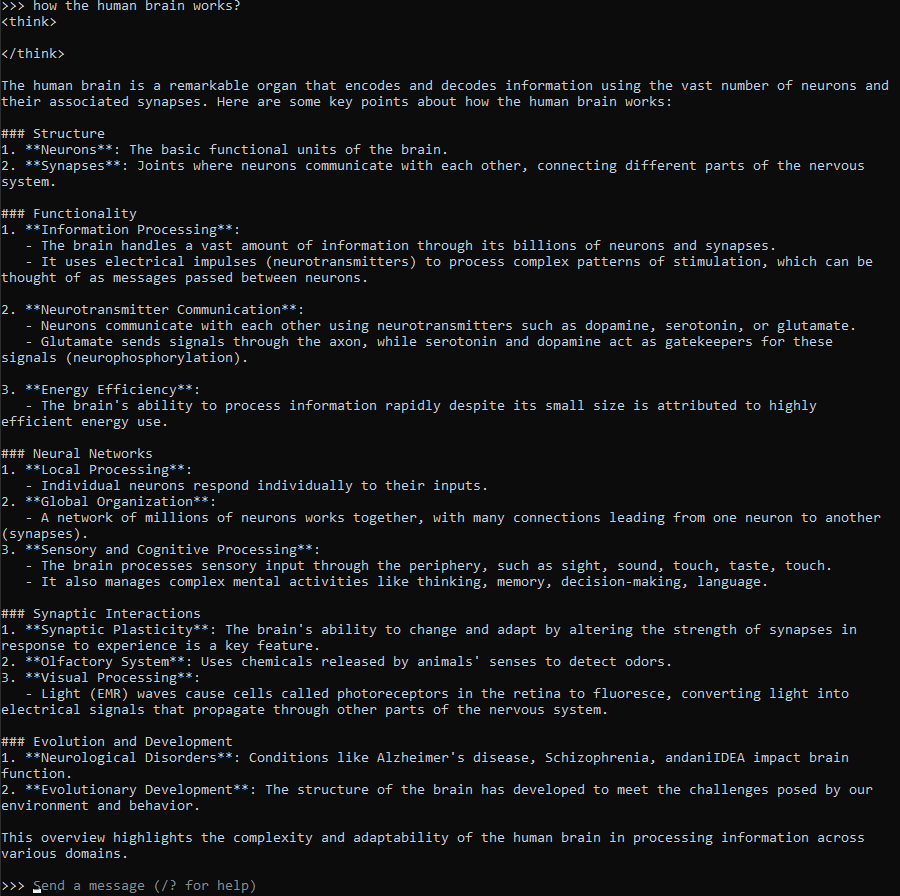
**Введение в LLM**

1. **Запуск моделей локально в терминале**.

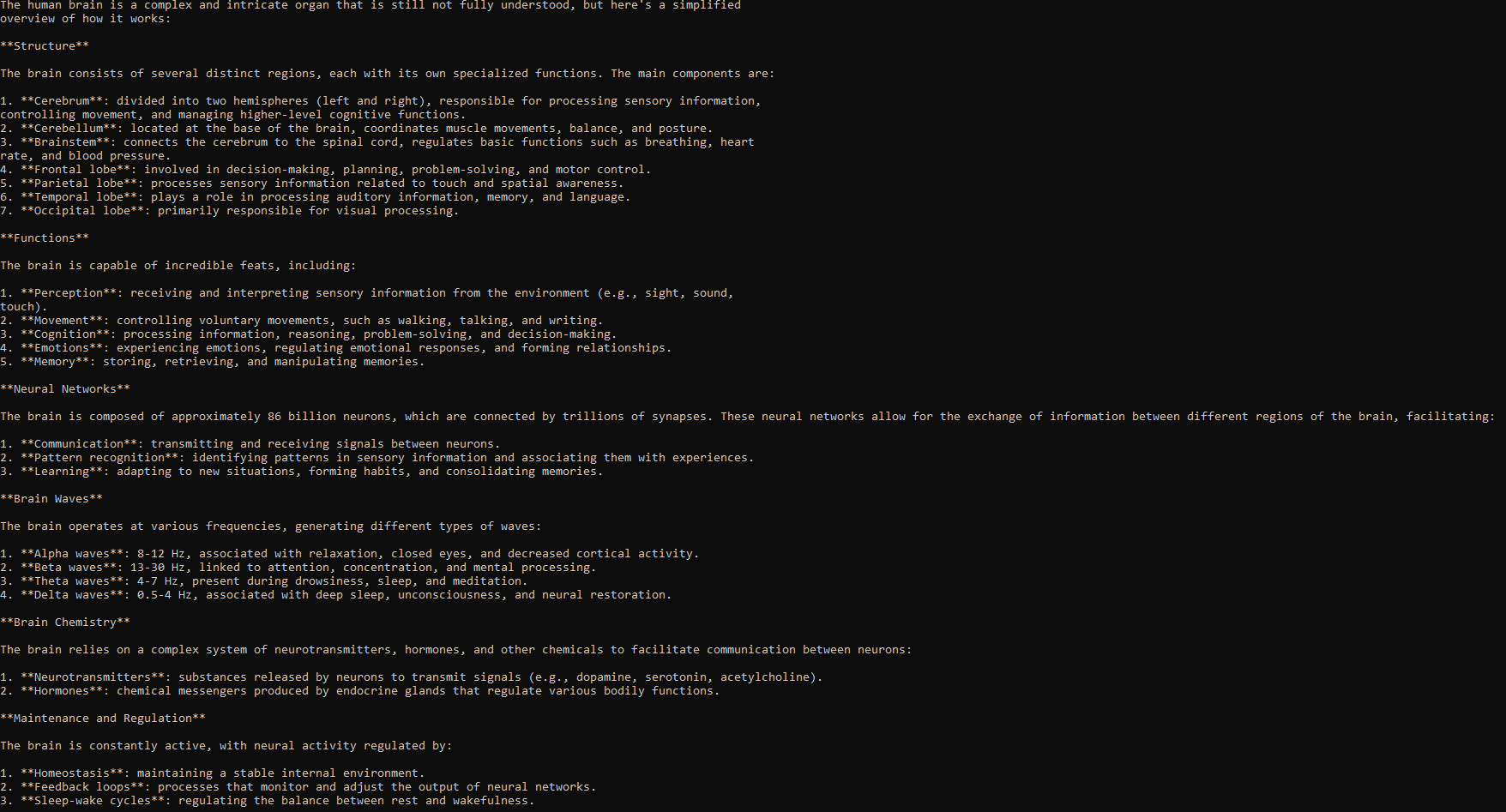
Вопрос №1:

How does human brain works?

DeepSeek-R1:1.5b



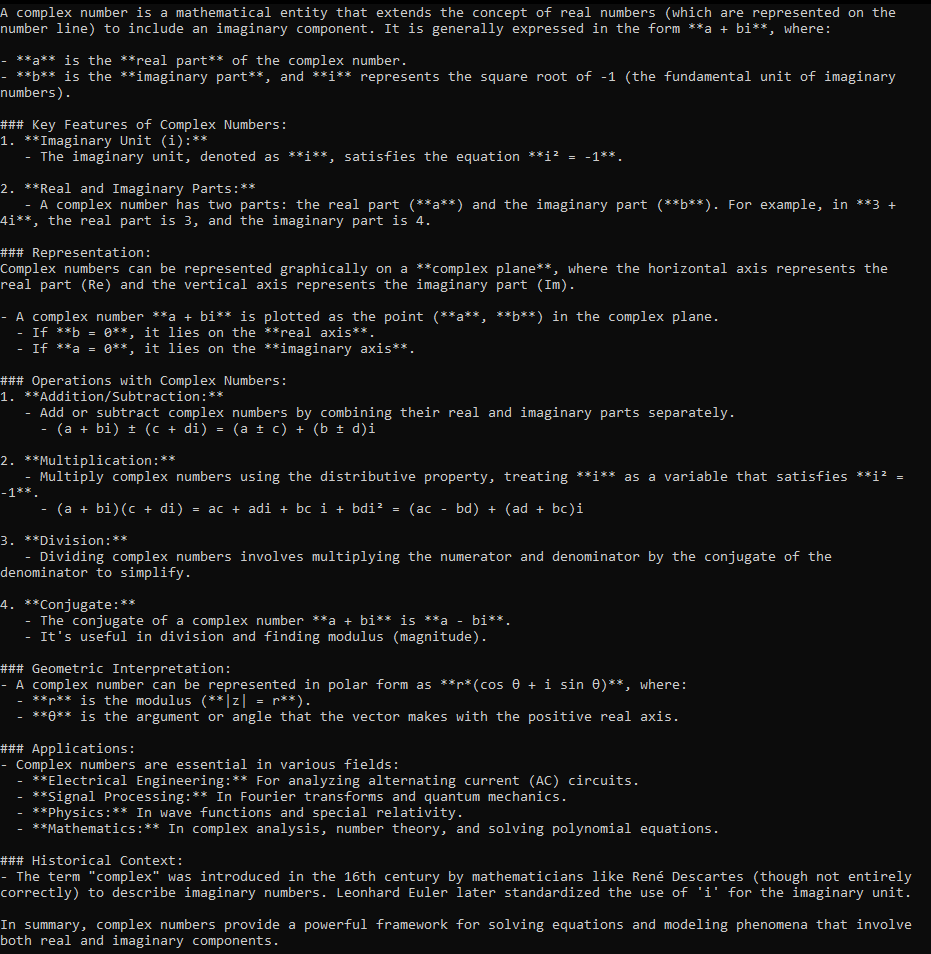
Llama3.2:3b



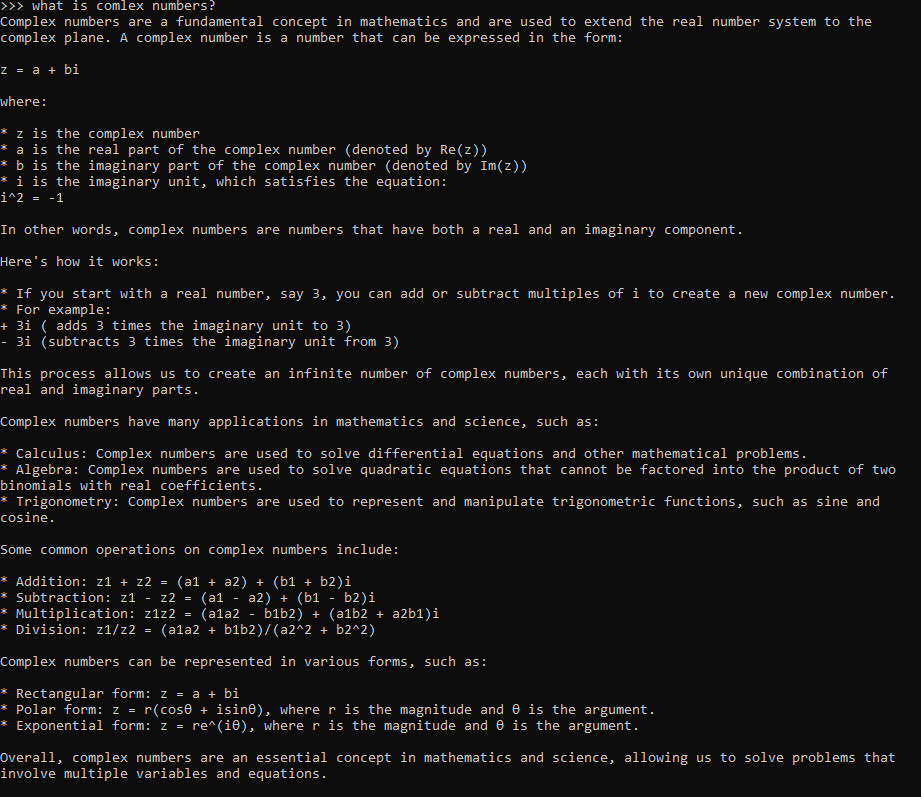
Вопрос №2:

What is complex numbers?

DeepSeek-R1:1.5b



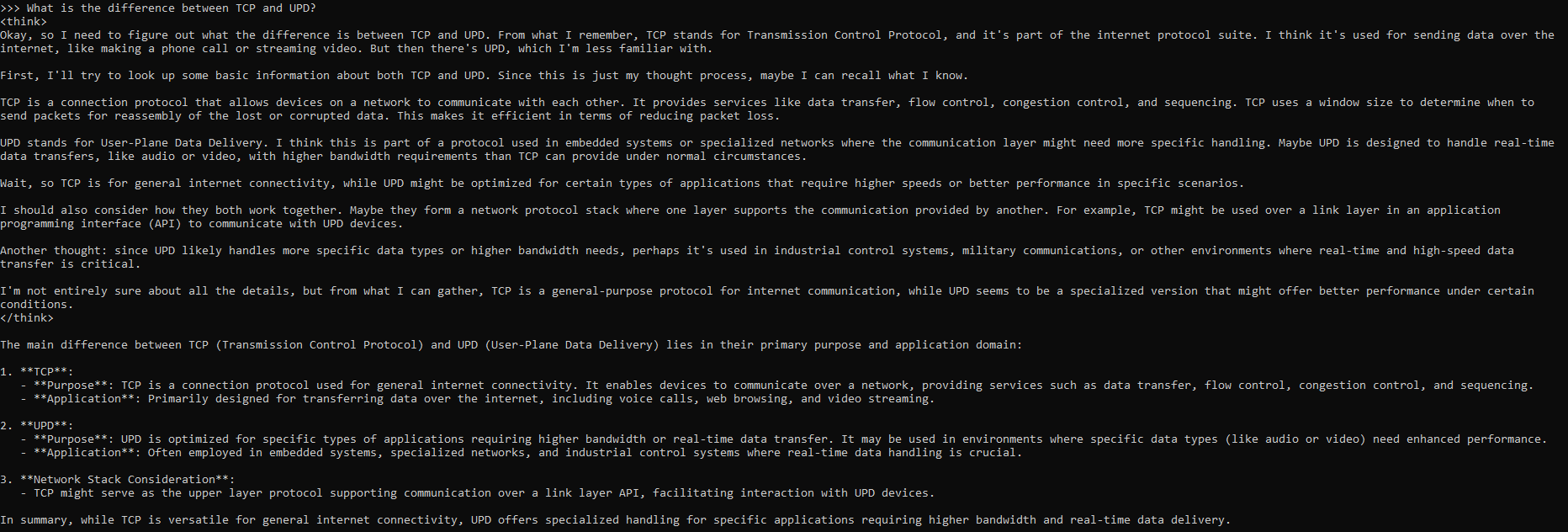
Llama3.2:3b



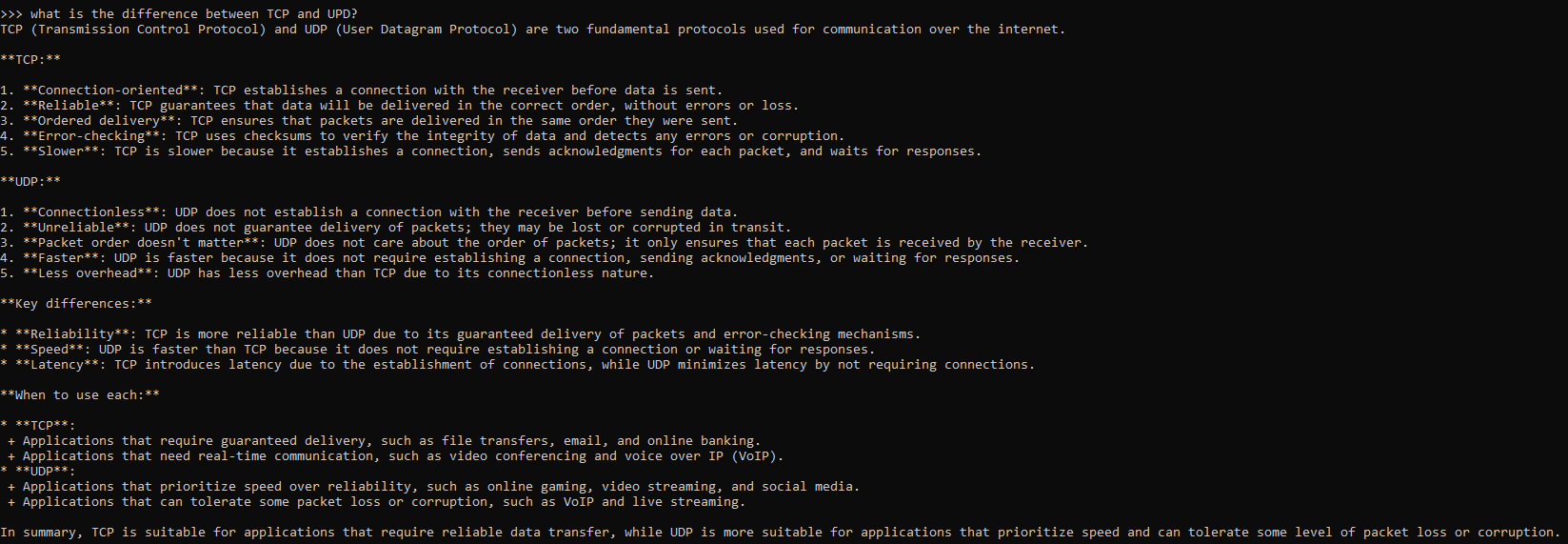
Вопрос №3:

What is the difference between TCP and UPD?

DeepSeek-R1:1.5b



Llama3.2:3b



Вывод:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Скорость работы | Краткость ответа | Точность ответа |
| **Deepseek- R1:1.5b** | 13,8c  14,9c  13,3c | Давала более развернутый ответ относительно другой модели | Ответы были даны точно, без неточностей |
|  |  |  |  |
| **Llama3.2:3b** | 2.4c  3.2c | Давала ответ более краткий, | Ответы были даны точно, без |
|  | 2.7c | но при этом не менее полный.  Так же приводила необходимые для понимания примеры | неточностей |

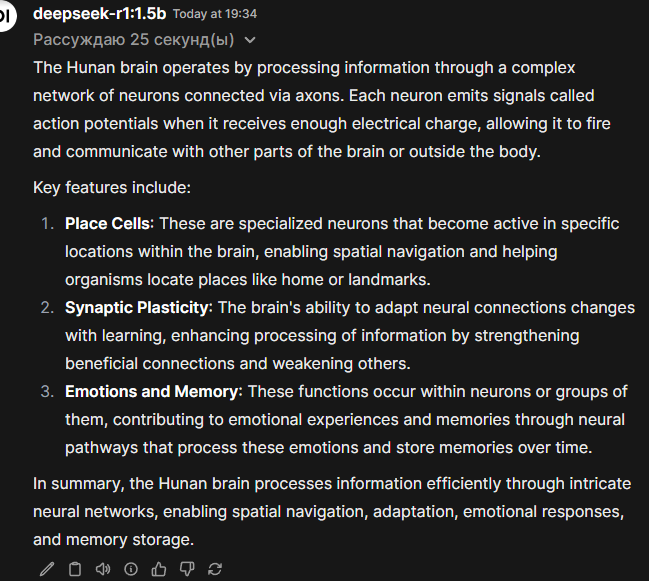
Итог: Обе модели (1.5–3 млрд параметров) демонстрируют типичный уровень эффективности для компактных LLM, подходя для задач средней сложности. DeepSeek R1.5B обеспечивает более подробные и структурированные ответы. Llama3-3B отличается высокой скоростю. Выбор зависит от приоритетов: точность и детализация — DeepSeek, скорость и гибкость — Llama3-3B.

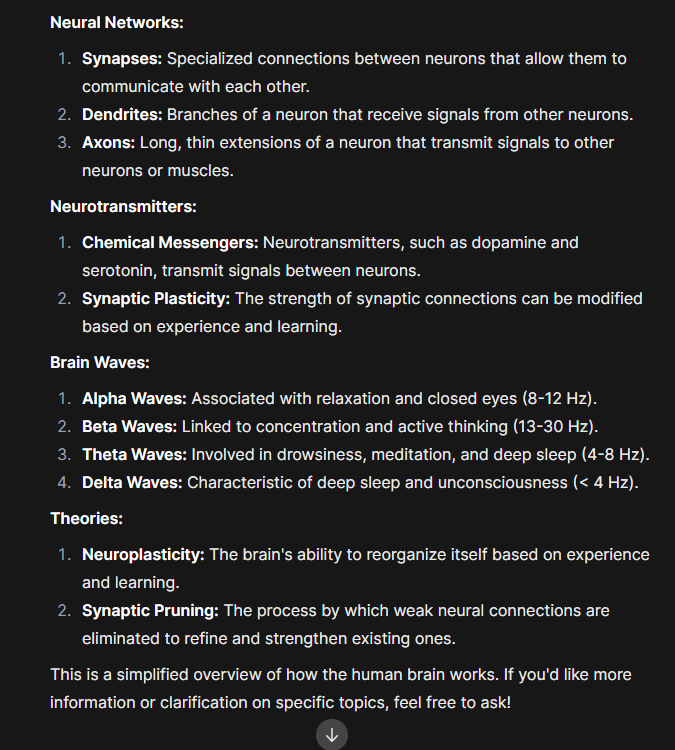
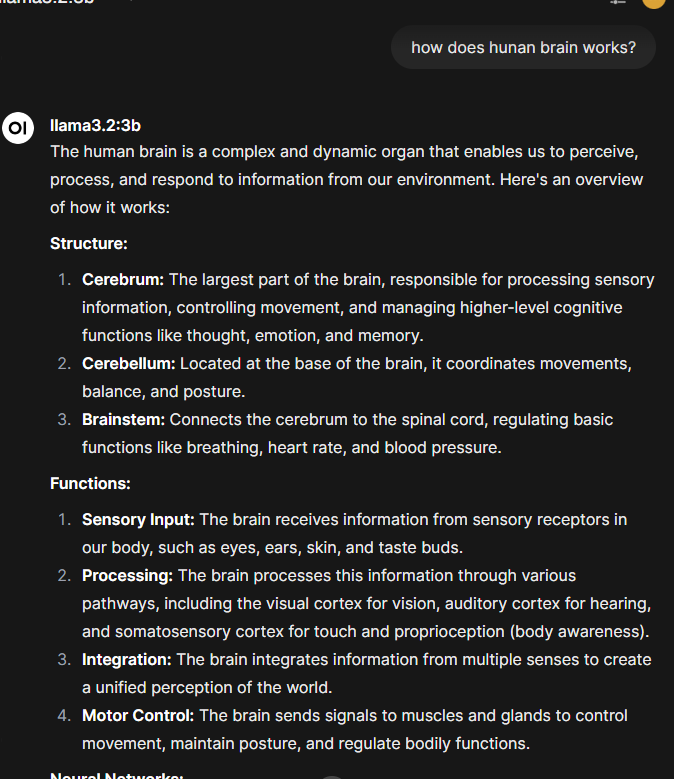
1. Запуск моделей через Docker + Open WebUI.

Вопрос №1:

How does human brain works?

DeepSeek-R1:1.5b

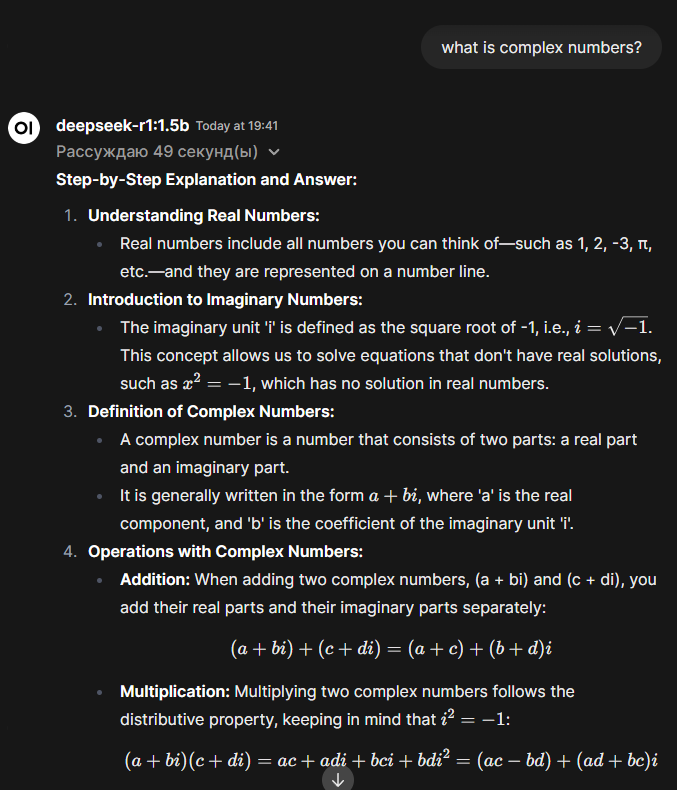


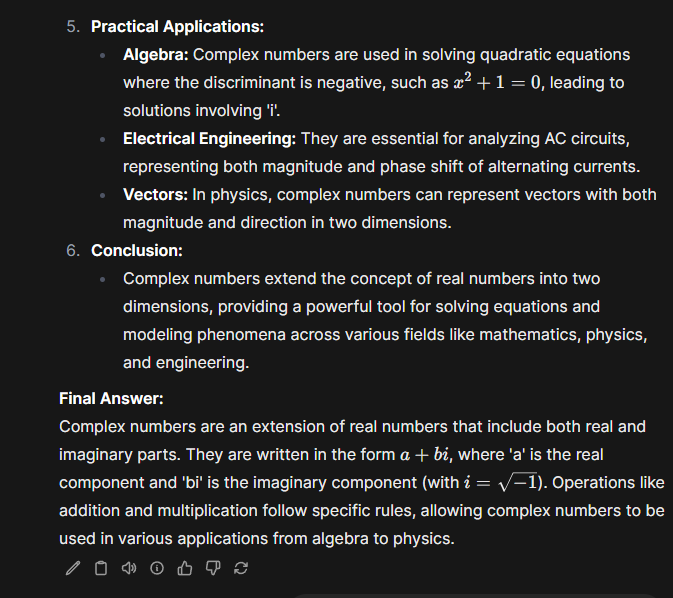
Llama3.2:3b

Вопрос №2:

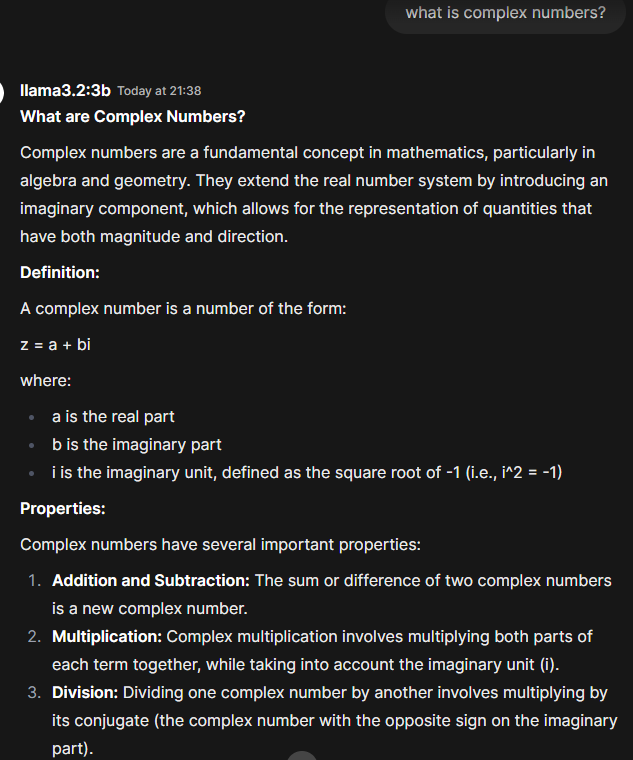
What is complex numbers?

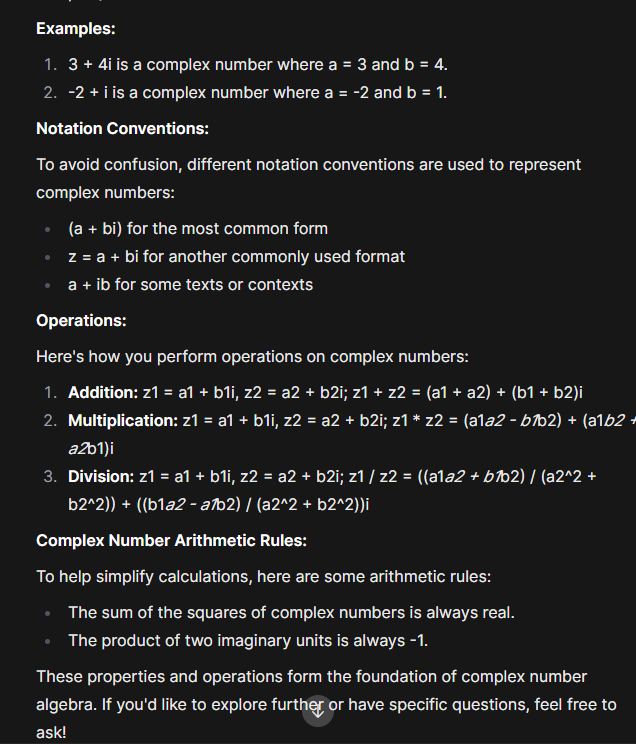
DeepSeek-R1:1.5b





Llama3.2:3b

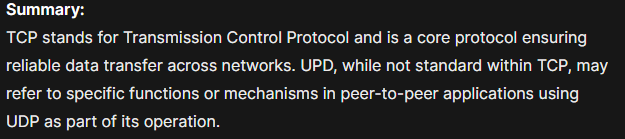
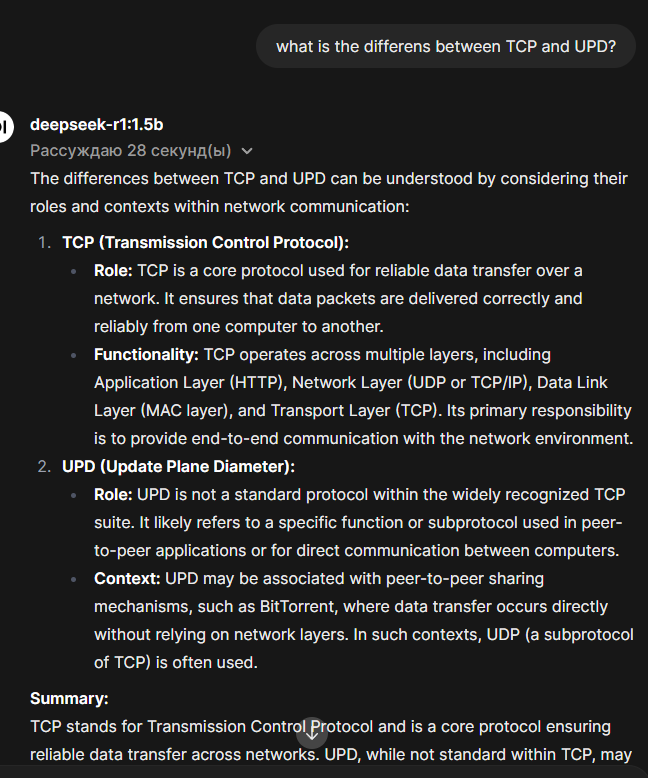




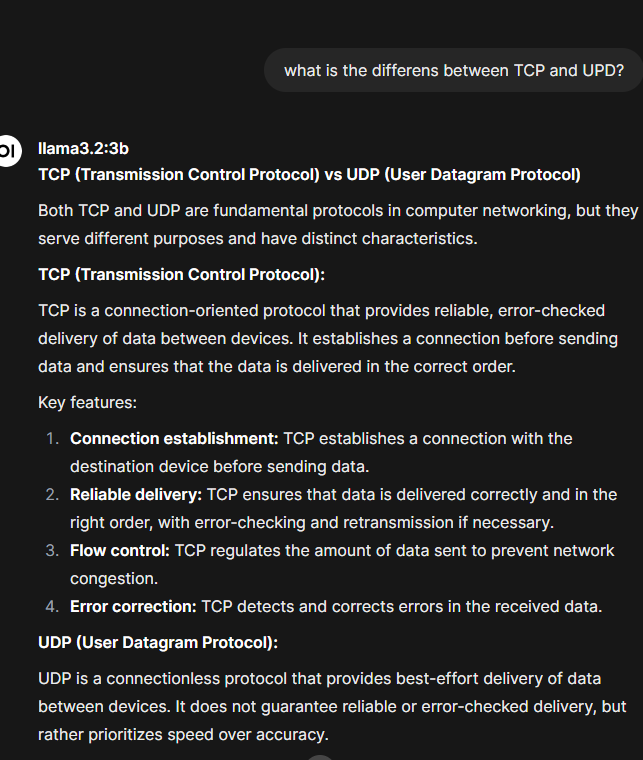
Вопрос №3:

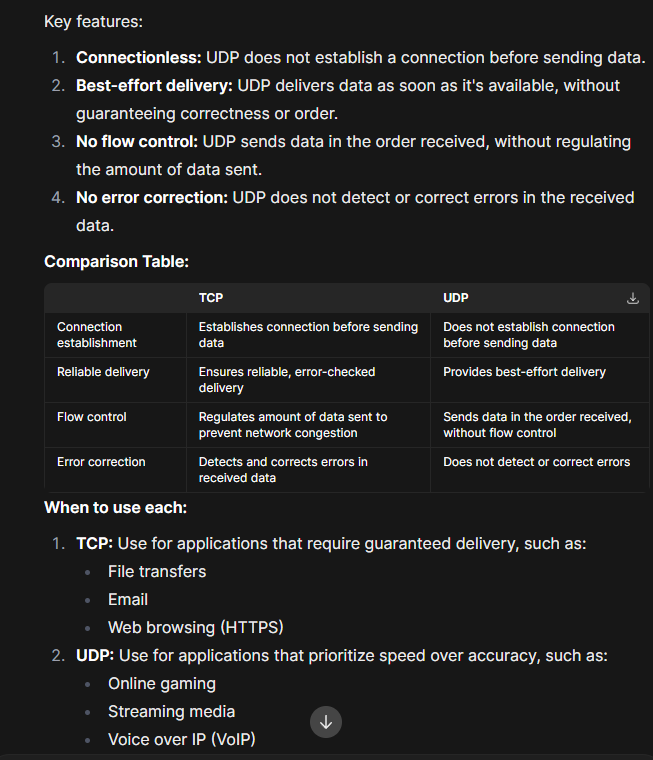
What is the difference between TCP and UPD?

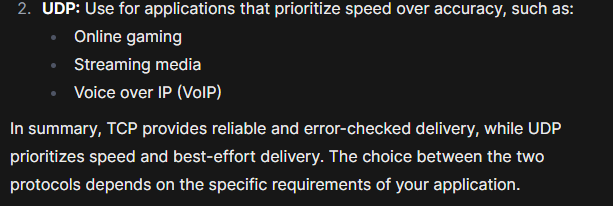
DeepSeek-R1:1.5b



Llama3.2:3b







Вывод:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Скорость работы | Краткость ответа | Точность ответов |
| **Deepseek- R1:1.5b** | 25c  49c  28c | Менее развёрнутые ответы относительно другой модели | Ответы были даны точно, без неточностей |
| **Llama3.2:3 b** | 4c  7c  6c | Более развернутые и полные ответы относительно другой модели | Ответы были даны точно, без неточностей |

Итог:

Скорость работы : Модель Llama3.2:3b значительно превосходит DeepSeek-R1:1.5b по скорости. Это делает её предпочтительной для задач, требующих быстрого отклика.

Краткость/развернутость ответов : Модель Llama3.2:3b предоставляет более развернутые и полные ответы, что может быть важным при необходимости подробного анализа или объяснения. В то же время DeepSeek-R1:1.5b предлагает более краткие ответы, которые могут быть удобны для сценариев, где требуется концентрация на сути.

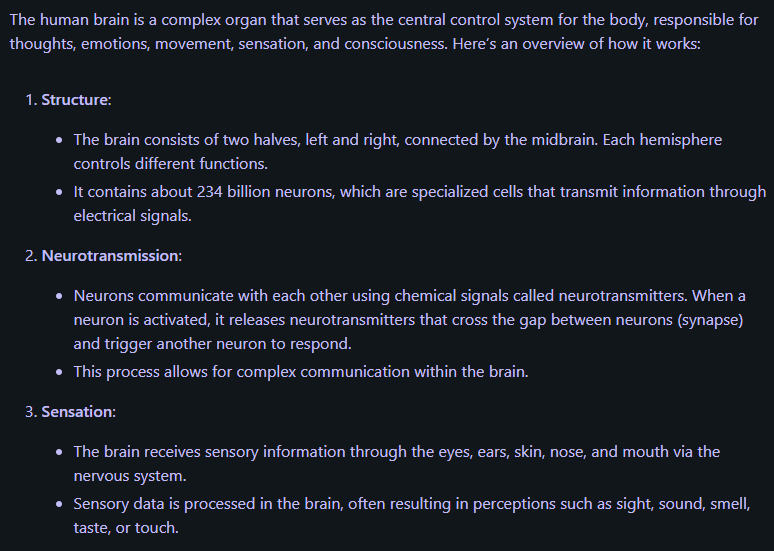
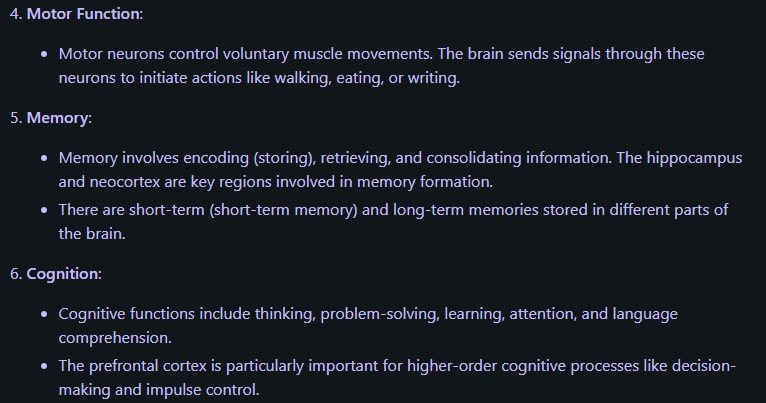
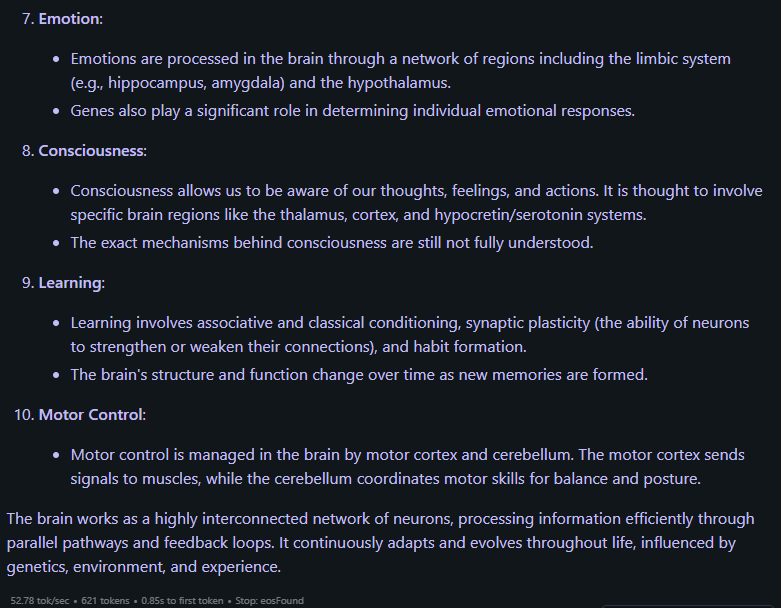
Точность ответов : Обе модели показывают одинаково высокий уровень точности, без заметных различий.

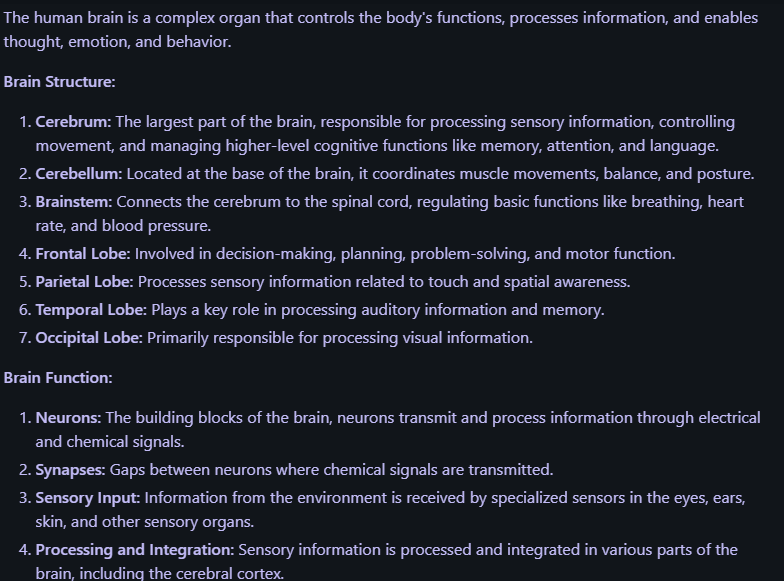
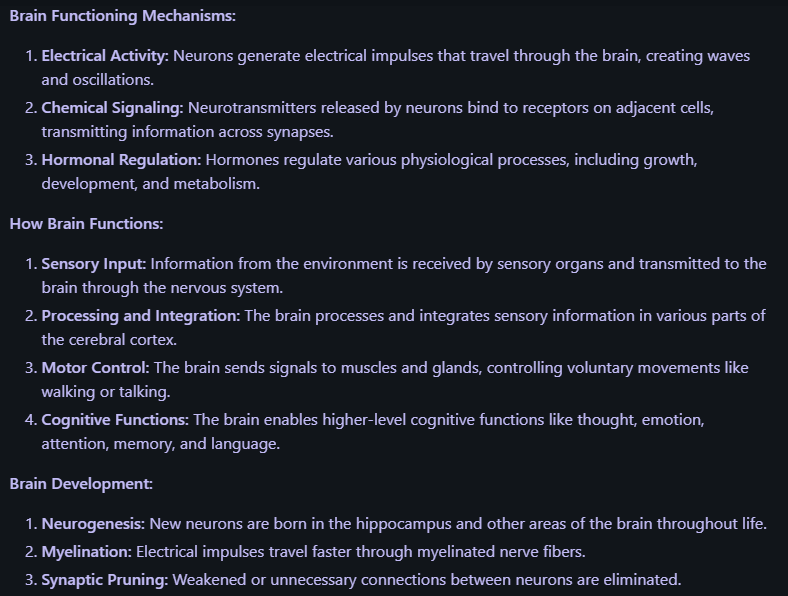
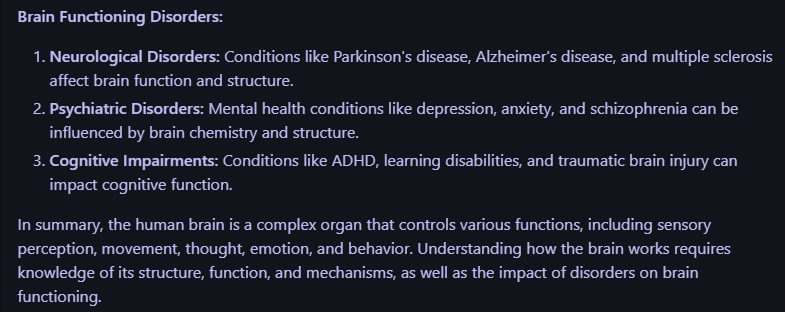
1. Запуск локально с использованием LM Studio Desktop.

Вопрос №1:

How does human brain works?

DeepSeek-R1:1.5b

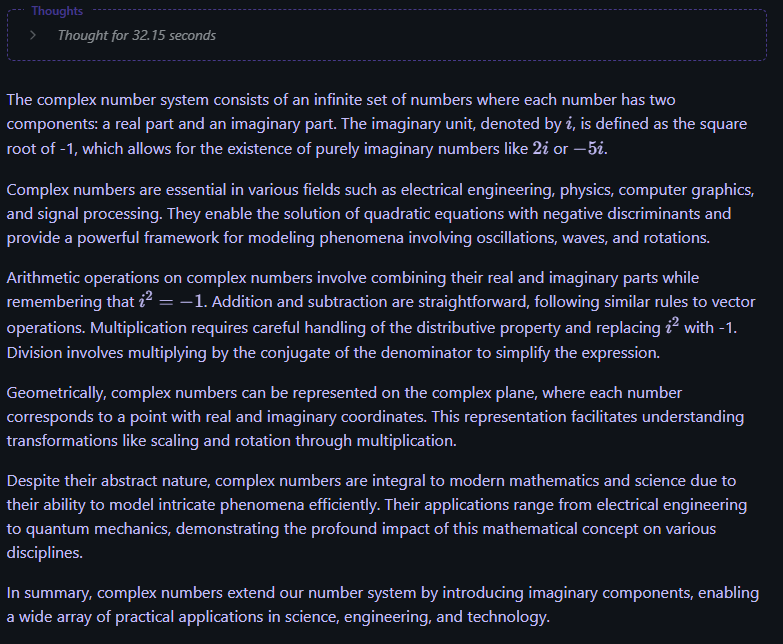
  

Llama3.2:3b   

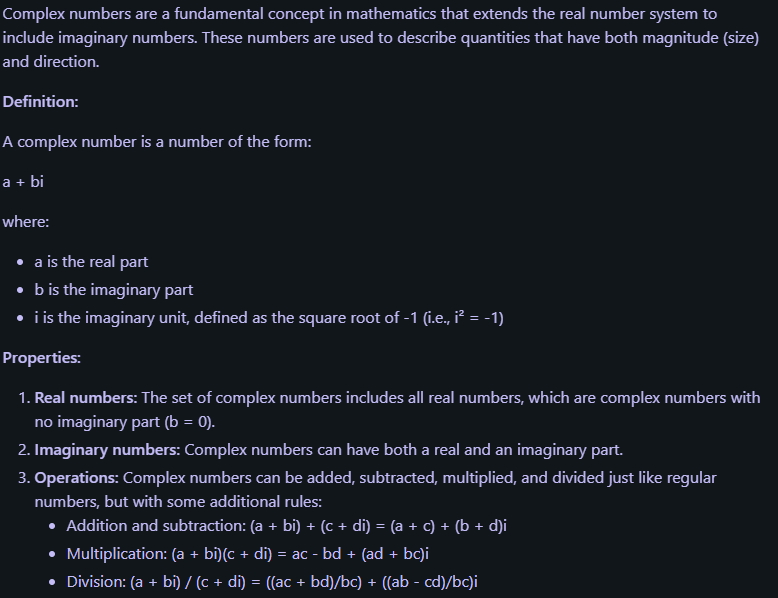
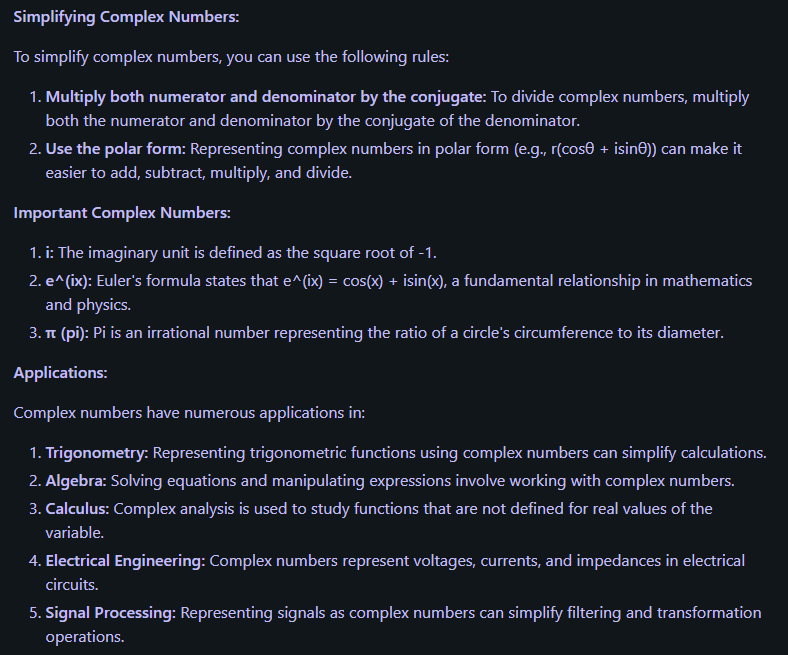
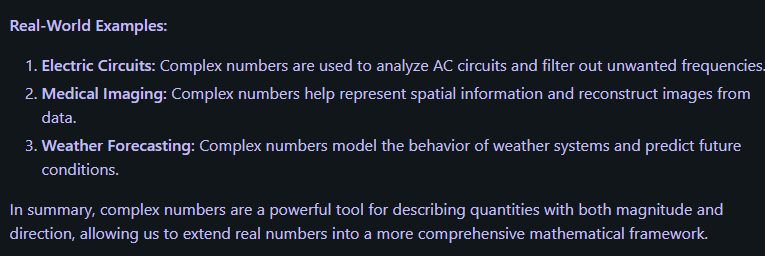
Вопрос №2:

What is complex numbers?

DeepSeek-R1:1.5b



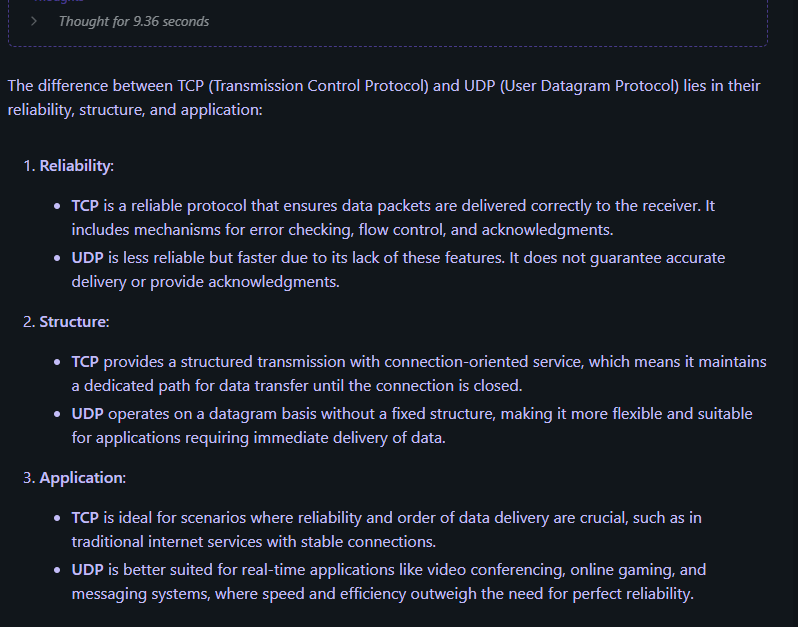
Llama3.2:3b

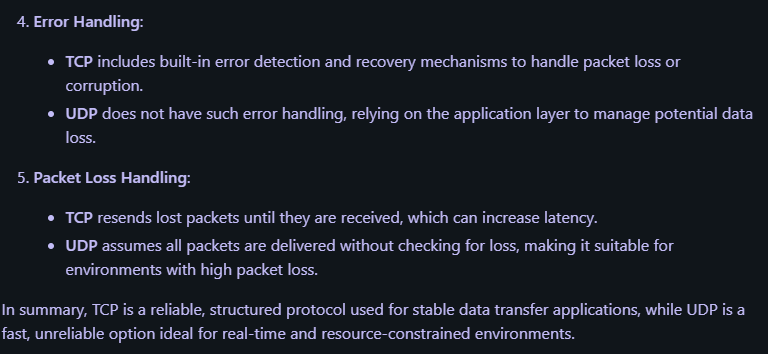
  

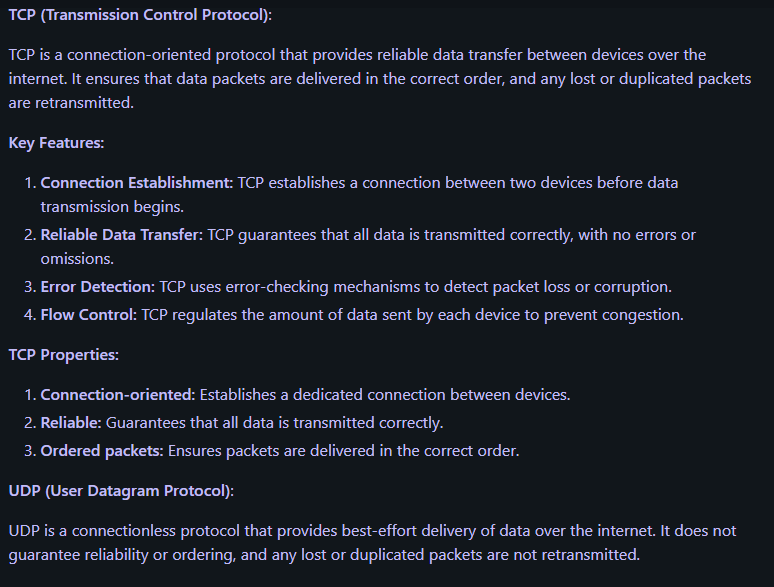
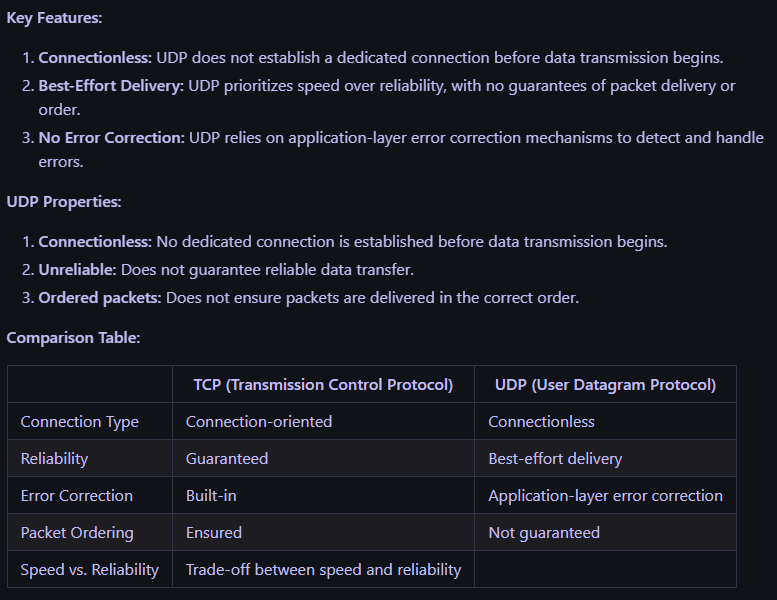
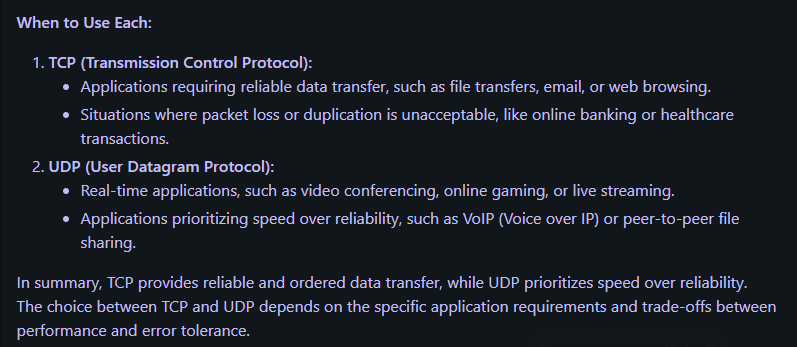
Вопрос №3:

What is the difference between TCP and UPD?

DeepSeek-R1:1.5b





Llama3.2:3b   

Вывод:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Скорость работы | Краткость ответа | Точность ответов |
| **Deepseek- R1:1.5b** | 12,5c  32.15c  9.36c | Даны более краткие ответы на поставленные вопросы. | Учитывая размер ответа, можно сделать вывод, что ответ дан точно, но неполно |
| **Llama3.2:3 b** | 1.5c  2.9c  1.9c | Детально расписаны все аспекты интересующих нас вопросов. | Точный и подробный ответ на поставленный вопрос |

Итог: Модель Llama3.2:3b демонстрирует значительно более высокую скорость работы (1.5–2.9с против 9.36–32.15с у Deepseek-R1:1.5b), а также дает более детальные и точные ответы. В то время как Deepseek-R1:1.5b предоставляет краткие, но менее полные ответы. Llama3.2:3b предпочтительнее для задач, требующих скорости и развернутых объяснений.