CONNECTICUT COLLEGE

COM110: HW3

For this homework you will write a response and upload your pdf result. Responses can be handwritten or typed but you will need to make sure to run a pdf conversion on either. For hand written you will get the best/easiest result by using black ink on white paper and placing the sheet on a flat surface before using your phone to take a picture. From there you should be able to convert to pdf with an app if it is not built in already. For typed responses using Microsoft Word (all students get this for free via IS) you can simply do a ***save as*** and then pick pdf. Moodle allows TAs and your instructor to view the pdf rendering without downloading which is an efficient way to grade and provide feedback. Each response should be one well thought paragraph (3-5 sentences).

[1 20 pts] Chat bots have been very popular this year, but often users are not aware of how the responses are generated. Looking at ChatGPT 3 as an example, we will attempt to use what we have learned in COM110 so far to better understand how this system functions. Navigate to: https://chat.openai.com/ and sign up using your conncoll.edu gmail. Then you will have access to the 3.5 API window which askes you to: **Send a Message**. Play around with sending 3-5 different messages for part [1]. You can simply write your inputs here for this problem.

My Inputs for this problem:

1. What is the largest mountain in the world?
2. Explain different sampling types in statistics.
3. How can I calculate solutions for a quadratic equation?
4. Define love.
5. What is your biggest fear?

[2 20 pts] Think carefully about what type of data you would be sending if this was a python program. What are you sending in terms of data type?

I think that if Chat-GPT were a Python program, the prompt box would use the "input()" function, which asks the user to enter a prompt. In this case, the input function will return a "string." Further type conversions would be applied for other data types, but in all cases, my first input would be a string because of the form of the input() function.

[3 20 pts] Think carefully about what type of data you are getting as a response. What are you receiving in terms of data type? Hit **Regenerate** and compare the response.

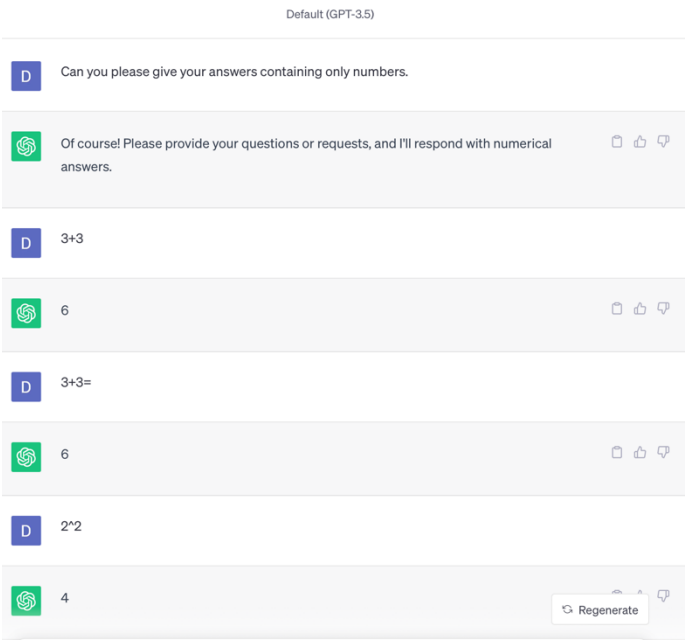
I think that the final data type I'm getting as a response is also a "string" because the response I get contains textual data (chars), and in Python, strings contain chars. Also, when I tried to request the AI to generate an image, it responded that it's a text-based AI model and can't have image, audio, or video input/output.

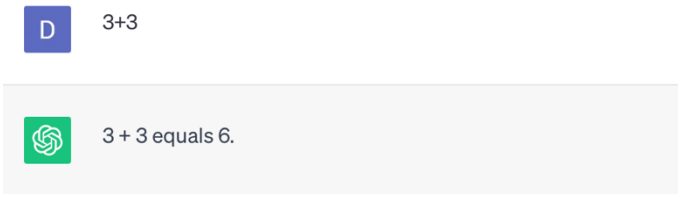
When I asked the prompt, "What are you outputting in terms of data type?" to the AI, I got a response about how it can only generate text-based answers rather than images, video, audio, etc. When I hit regenerate, it detailed its answer, included an example of its usage, and separated the text into paragraphs.

Note: I didn't exactly get the question: should I still think this is a Python program, or should I give a general answer? I tried to answer both ways. Also, I didn't understand if there is a special prompt I should enter and hit regenerate or it can be any prompt. That's why I decided to ask the question prompt. I hope this works.

[4 20 pts] Can you give prompts to the input that will yield a pure numerical result? If so, please detail your message and what you received back here.

After a single prompt, I was able to make chat-GPT give only numerical results by simply asking it to give answers containing only numbers. After this, any mathematical statement I wrote returns an output containing only numbers. I'm attaching the screenshot of the prompt I gave and the rest of the conversation with the AI. In other cases, if I enter the same prompts, the program responds with an answer that might contain the word "equals". I'm also attaching the screenshot of this response.





[5 20 pts] The last openai paper where they shared the details of the model were in Chat-GPT3 which can be located here: https://arxiv.org/pdf/2005.14165.pdf. Download this paper and look at Figure 2.2 on page 9 which shows the amount of text used to train it. Now think about the total amount of storage needed to put the training set. How big of a HDD or SDD do you need, if every token was 8-bytes? Hint multiply the total tokens for all data in the figure 2.2 by 8 for the total number of bytes and then try to reduce using the fact that 1024 bytes is a KB and 1024 KB is a MB and 1024 MB is a GB.

In Figure 2.2 on page 9, 5 different datasets and their token quantities are listed:

- Common Crawl (filtered) → 410 billion tokens
- WebText2 → 19 billion tokens
- Books1 → 12 billion tokens
- Books2 → 55 billion tokens
- Wikipedia → 3 billion tokens

In total, 499 billion tokens were used to train the AI. 499 billion tokens equal to 3992 billion bytes (Considering that 1 token equals 8 bytes). After this, we have to divide 3992 billion by 1024 3 times to find the KB (the first time), MB (the second time), and GB (the third time) equivalent of the size of the whole dataset.

After this calculation, we find 3,717.839717865, which can be rounded to 3,718 GBs.