AI for Research:
Writing Effective Prompts for Coding with LLMs

Aaron Geller
(with help from Emilio Lehoucq)
Research Computing and Data Services

Northwestern | INFORMATION TECHNOLOGY

This workshop is brought to you by:

Northwestern IT
Research Computing and Data Services

Need help?

- AI, Machine Learning, Data Science
- Statistics
- Visualization
- Collecting web data (scraping, APIs), text analysis, extracting information from text
- Cleaning, transforming, reformatting, and wrangling data
- Automating repetitive research tasks
- Research reproducibility and replicability
- Programming, computing, data management, etc.
- R, Python, SQL, MATLAB, Stata, SPSS, SAS, etc.

Request a **FREE** consultation at **bit.ly/rcdsconsult**.

Upcoming RCDS Workshops

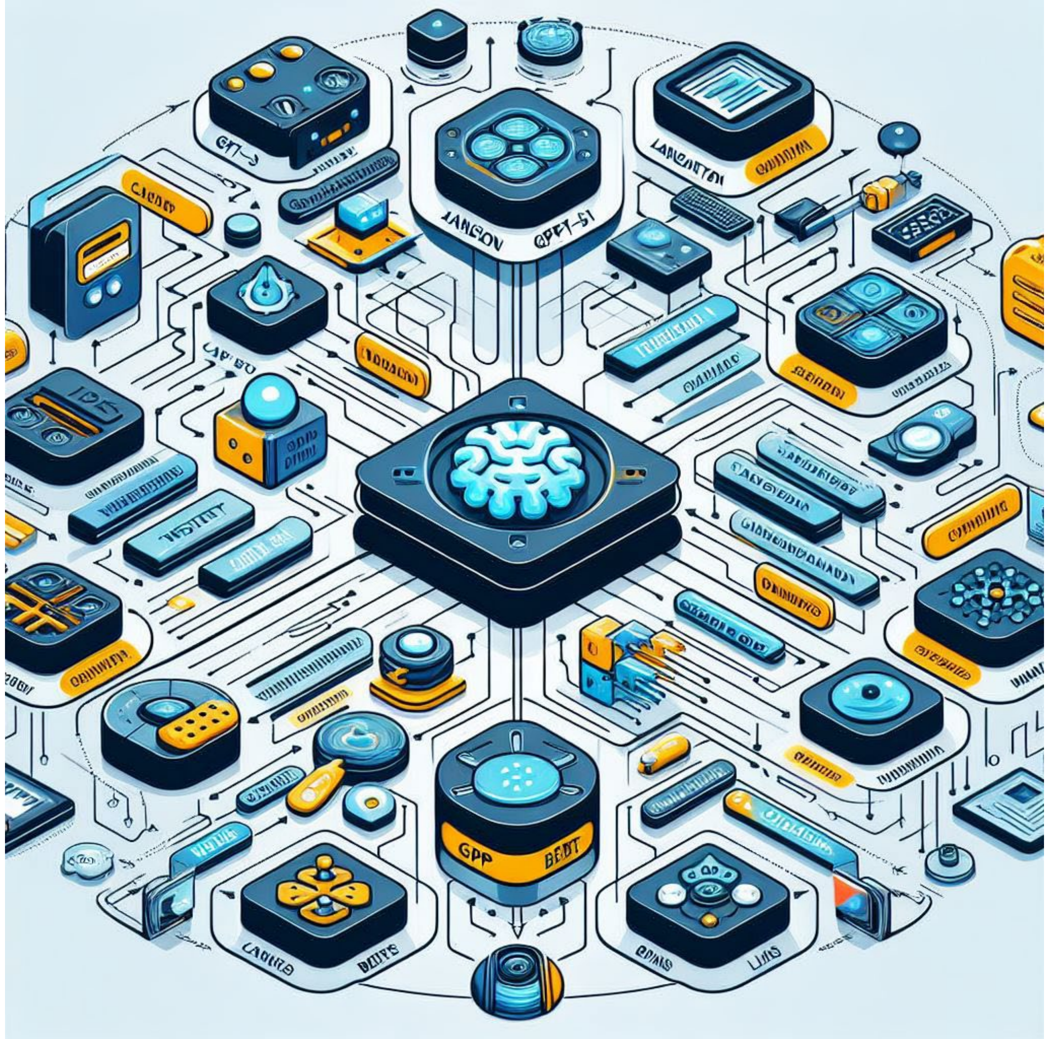
https://www.it.northwestern.edu/departments/it-services-support/research/research-events.html

Prompt Engineering

How to write effective prompts for LLMs

What Can You Ask LLMs To Do When Coding?

- Brainstorm ideas, pros/cons
- Initial structure or template for code
- **Generate code**
- Generate unit tests
- Simplify or refactor code
- Debug
- Explain code that you wrote or found



^ Response from DALL-E 3 for "Please generate an image about the many possible uses of ChatGPT and LLMs."

How To Write Prompts for Coding
(1)

Clear, detailed, non-ambiguous, and concise.

DO	***DON'T***
I have 3000 Reddit posts. I want to use topic modeling to find out what the posts are talking about. Before doing the topic modeling, how do I need to prepare the text?	How to pre-process text for NLP?

How To Write Prompts for Coding
(2)

Tell the LLM what role to take, and your own expertise level.

DO	***DON'T***
You are an expert in computational biology, particularly genomics. What are the key R packages for statistics?	What are the key R packages for statistics?

How To Write Prompts for Coding

(3)

Describe the programming language, libraries, and other technologies.

DO	***DON'T***
How do I scrape a website in Python using Selenium?	How do I scrape a website in Python?

How To Write Prompts for Coding

(4)

Explain what the code is for.

DO	***DON'T***
I want to scrape a website that relies heavily on JavaScript. I want to get the source code after loading all elements. How can I do that with Python?	I want to scrape a website. How can I do that with Python?

How To Write Prompts for Coding

(5)

Specify any constraints or requirements.

DO	*DON'T*
Using ggplot2 in R, I want to iterate over 100 columns, creating the same plot with each. Please write code to do that considering that I'm running the code in my university's high-performance computing cluster.	Using ggplot2 in R, I want to iterate over 100 columns, creating the same plot with each. Please write code to do that.

How To Write Prompts for Coding

(6)

Be specific about what you want the code to do.

DO	***DON'T***
Write a function in Python that takes two numbers, adds them, and returns the result. The function should raise a type error if the inputs are not numbers.	Write a function in Python that takes two numbers, adds them, and returns the result.

How To Write Prompts for Coding

(7)

Provide examples of the desired behavior.

DO	***DON'T***
Write a regular expression to match salaries in job postings. These are some examples: String: Minimum $60,000 ANNUAL (12 months) Match: $60,000 String: Rate of Pay $78,035 - $106,517 Match: $78,035 and $106,517 String: starting salary will be from £46,047 up to £61,823 Match: £46,047 and £61,823	Write a regular expression to match salaries in job postings.

How To Write Prompts for Coding

(8)

Ask the LLM to explain itself and any assumptions.

DO	***DON'T***
Write a regular expression to match salaries in job postings. Explain any assumptions that you're making about the data and what the match should be.	Write a regular expression to match salaries in job postings.

How To Write Prompts for Coding

(9)

Ask the LLM to work step-by-step.

DO	***DON'T***
What is this R code doing? iris %>% group_by(Species) %>% summarize_all(mean) %>% ungroup %>% gather(measure, value, -Species) %>% arrange(value) Work step-by-step.	What is this code doing? iris %>% group_by(Species) %>% summarize_all(mean) %>% ungroup %>% gather(measure, value, -Species) %>% arrange(value)

How To Write Prompts for Coding

(10)

Specify how you want the LLM to give you the information.

DO	***DON'T***
What are different packages for supervised learning in R, as well as their pros and cons? Provide the information in a table with only one row per package.	What are different packages for supervised learning in R, as well as their pros and cons?

How To Write Prompts for Coding
(11)

Reformulate prompts as needed!

Recap

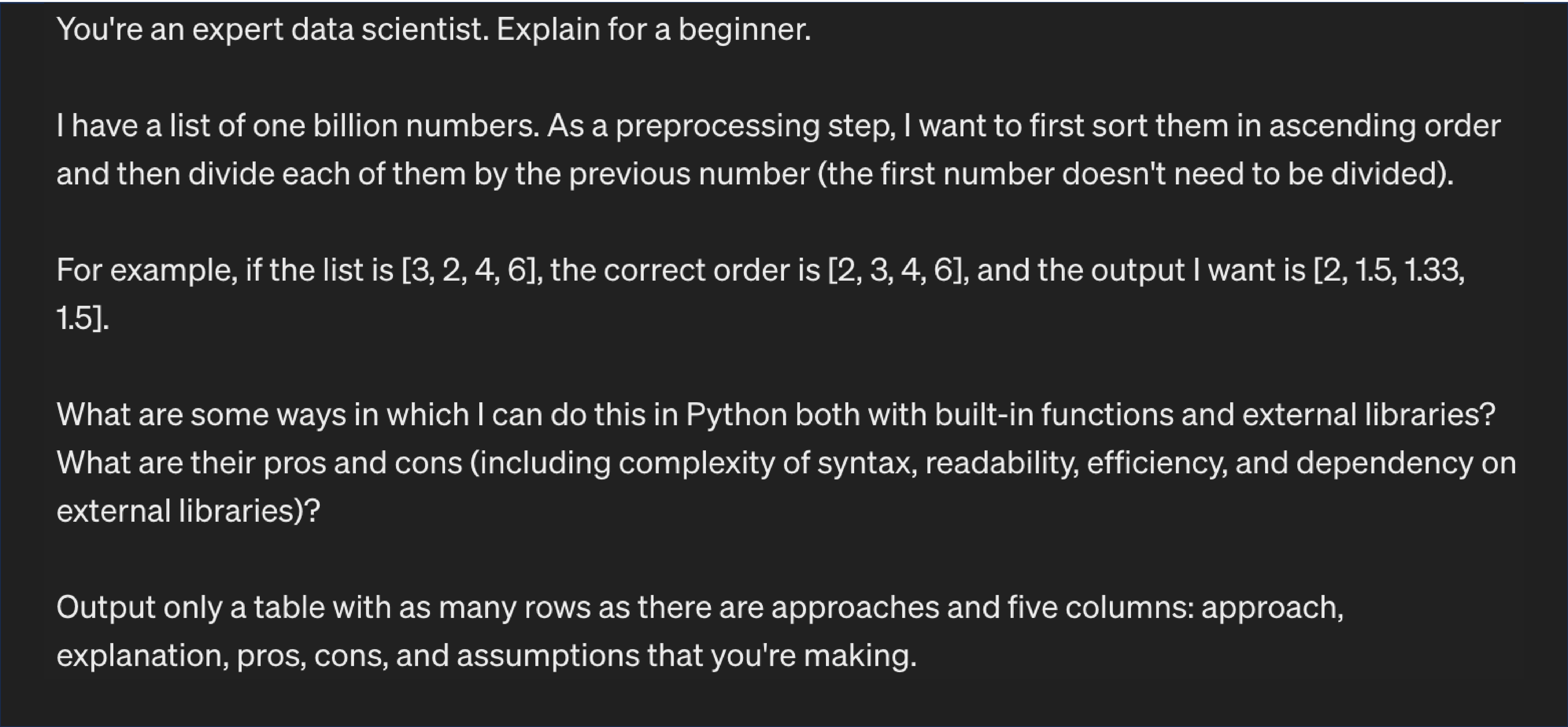
- Clear, detailed, non-ambiguous, and concise.
- Tell the LLM what role to take, and your own expertise level.
- Describe the programming language, libraries, and other technologies.
- Explain what the code is for.
- Specify any constraints or requirements.
- Be specific about what you want the code to do.
- Provide examples of the desired behavior.
- Ask the LLM to explain itself and any assumptions.
- Ask the LLM to work step-by-step.
- Specify how you want the LLM to give you the information.
- Reformulate prompts as needed!

How To Use Conversations For Coding

- LLMs are sensitive to chat history.
  - Treat interactions as conversations.
  - Use different conversations for different questions.
- Same prompt can give different results across conversations.
- You can copy code from one conversation to another to get feedback.
- Avoid long conversations.

^ Response from DALL-E 3 for "Please generate an image of a person having multiple conversations with a computer"

Example conversation



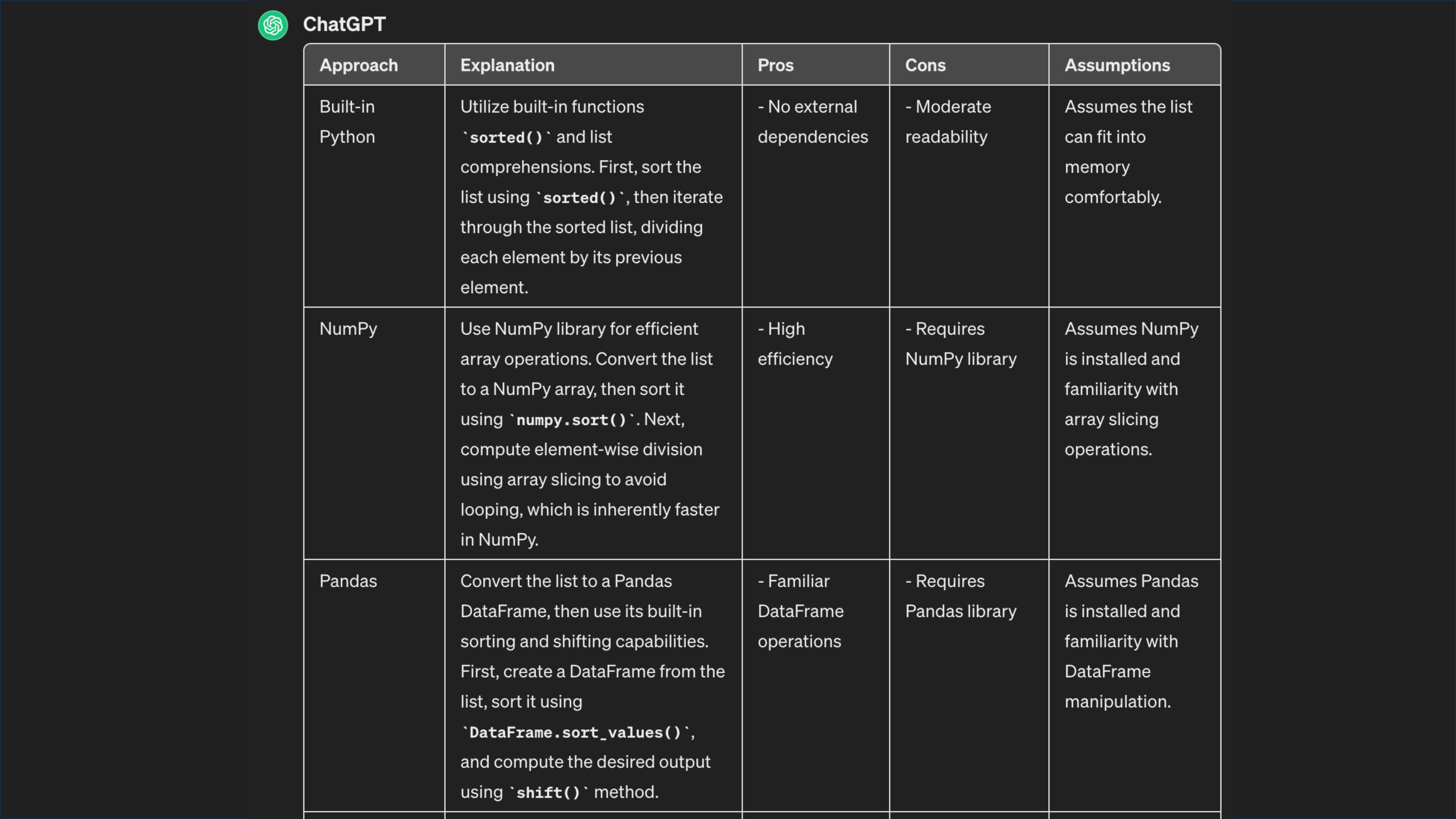
You're an expert data scientist. Explain for a beginner.

I have a list of one billion numbers. As a preprocessing step, I want to first sort them in ascending order and then divide each of them by the previous number (the first number doesn't need to be divided).

For example, if the list is [3, 2, 4, 6], the correct order is [2, 3, 4, 6], and the output I want is [2, 1.5, 1.33, 1.5].

What are some ways in which I can do this in Python both with built-in functions and external libraries? What are their pros and cons (including complexity of syntax, readability, efficiency, and dependency on external libraries)?

Output only a table with as many rows as there are approaches and five columns: approach, explanation, pros, cons, and assumptions that you're making.

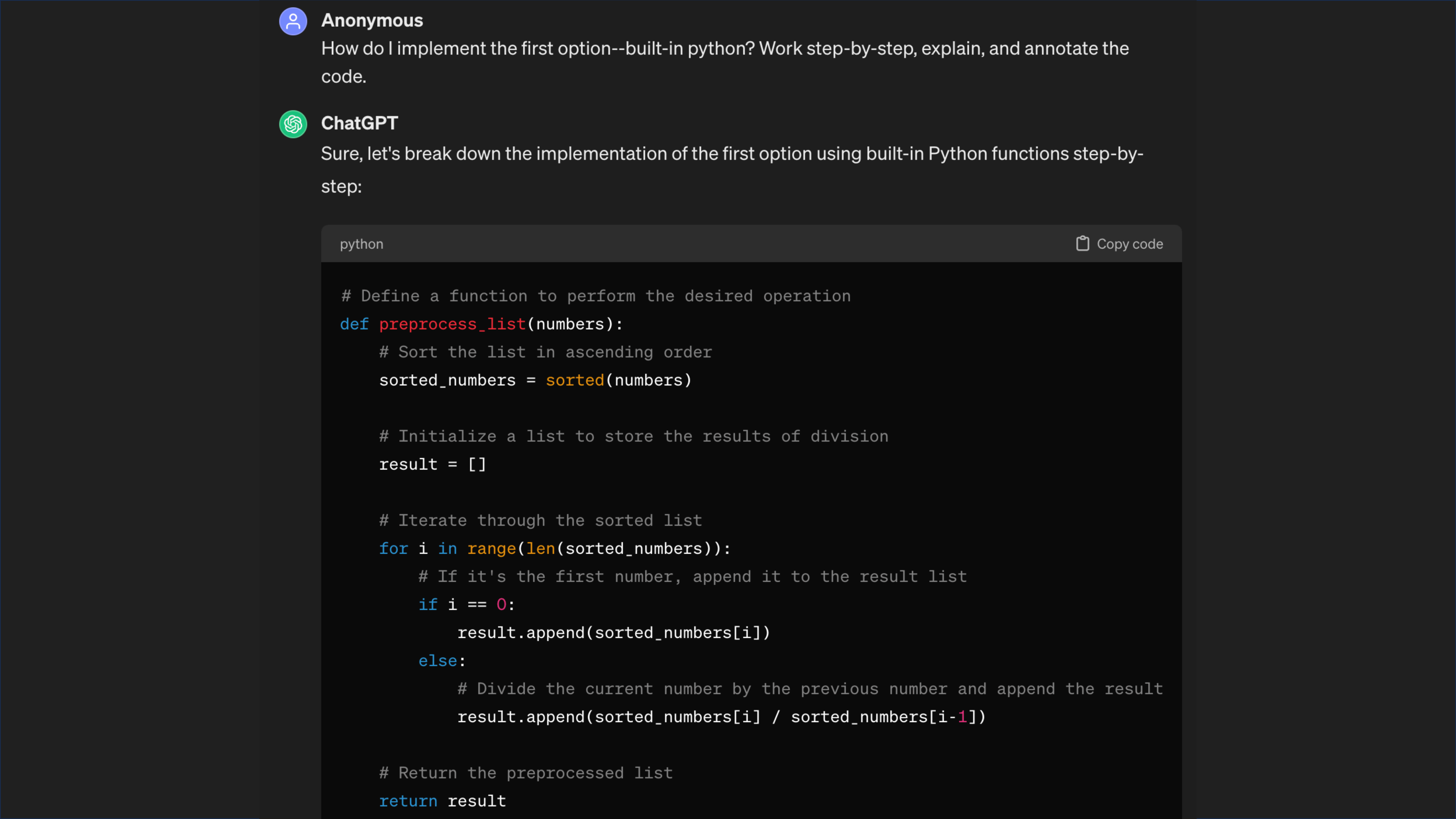


ChatGPT

Approach	Explanation	Pros	Cons	Assumptions
Built-in Python	Utilize built-in functions `` `sorted()` `` and list comprehensions. First, sort the list using `` `sorted()` ``, then iterate through the sorted list, dividing each element by its previous element.	- No external dependencies	- Moderate readability	Assumes the list can fit into memory comfortably.
NumPy	Use NumPy library for efficient array operations. Convert the list to a NumPy array, then sort it using `` `numpy.sort()` ``. Next, compute element-wise division using array slicing to avoid looping, which is inherently faster in NumPy.	- High efficiency	- Requires NumPy library	Assumes NumPy is installed and familiarity with array slicing operations.
Pandas	Convert the list to a Pandas DataFrame, then use its built-in sorting and shifting capabilities. First, create a DataFrame from the list, sort it using `` `DataFrame.sort_values()` ``, and compute the desired output using `` `shift()` `` method.	- Familiar DataFrame operations	- Requires Pandas library	Assumes Pandas is installed and familiarity with DataFrame manipulation.

Parallelization	Employ parallel processing to handle the sorting and division concurrently. Divide the list into chunks, sort each chunk independently, then perform division in parallel.	- Potential for significant speedup with large data	- Complexity in managing parallel processes	Assumes access to parallel processing libraries and familiarity with parallel programming concepts.
External Sort	Use an external sorting algorithm, like merge sort or external quicksort, suitable for large datasets that cannot fit into memory. After sorting externally, perform the division as usual.	- Suitable for datasets too large for memory	- Complexity in implementation	Assumes access to external sorting algorithms or libraries and familiarity with their implementation.

Each approach has its own trade-offs in terms of efficiency, readability, and dependencies. The choice depends on the specific requirements of the task, such as the size of the data, available computational resources, and familiarity with external libraries.



Anonymous

How do I implement the first option--built-in python? Work step-by-step, explain, and annotate the code.

ChatGPT

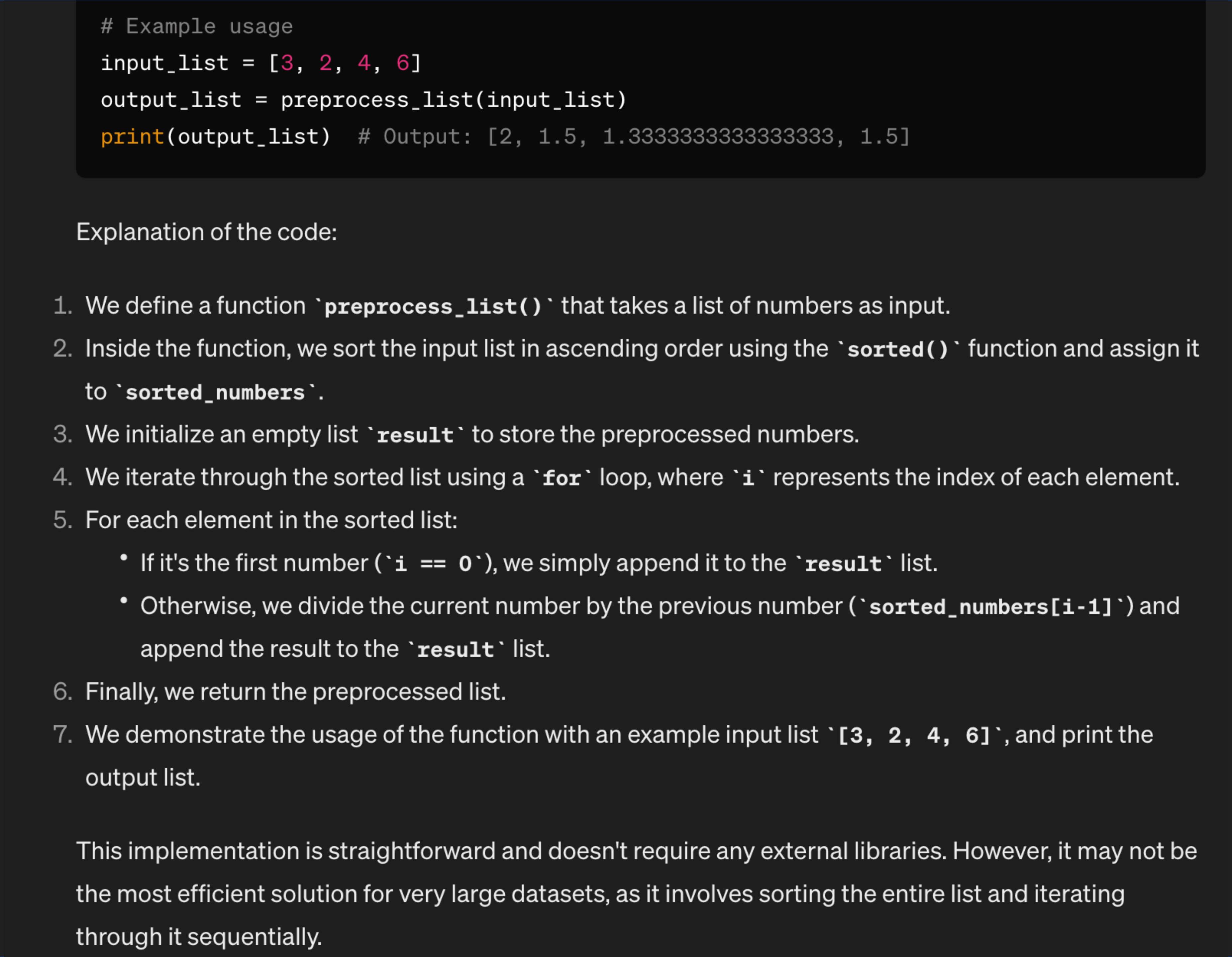
Sure, let's break down the implementation of the first option using built-in Python functions step-by-step:

```python
# Define a function to perform the desired operation
def preprocess_list(numbers):
    # Sort the list in ascending order
    sorted_numbers = sorted(numbers)

    # Initialize a list to store the results of division
    result = []

    # Iterate through the sorted list
    for i in range(len(sorted_numbers)):
        # If it's the first number, append it to the result list
        if i == 0:
            result.append(sorted_numbers[i])
        else:
            # Divide the current number by the previous number and append the result
            result.append(sorted_numbers[i] / sorted_numbers[i-1])

    # Return the preprocessed list
    return result
```

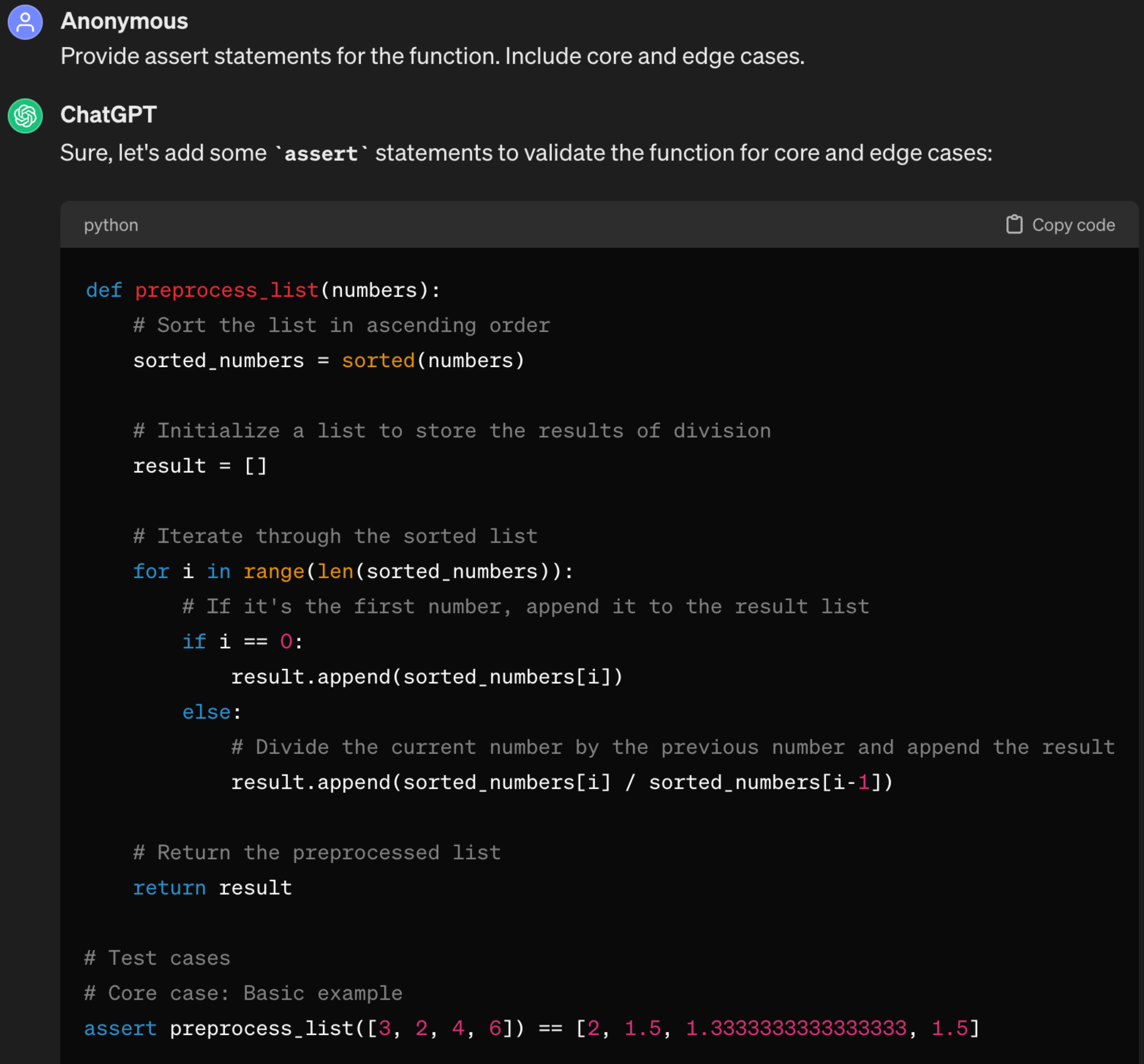


```
# Example usage
input_list = [3, 2, 4, 6]
output_list = preprocess_list(input_list)
print(output_list)  # Output: [2, 1.5, 1.3333333333333333, 1.5]
```

Explanation of the code:

1. We define a function `preprocess_list()` that takes a list of numbers as input.
2. Inside the function, we sort the input list in ascending order using the `sorted()` function and assign it to `sorted_numbers`.
3. We initialize an empty list `result` to store the preprocessed numbers.
4. We iterate through the sorted list using a `for` loop, where `i` represents the index of each element.
5. For each element in the sorted list:
   - If it's the first number (`i == 0`), we simply append it to the `result` list.
   - Otherwise, we divide the current number by the previous number (`sorted_numbers[i-1]`) and append the result to the `result` list.
6. Finally, we return the preprocessed list.
7. We demonstrate the usage of the function with an example input list `[3, 2, 4, 6]`, and print the output list.

This implementation is straightforward and doesn't require any external libraries. However, it may not be the most efficient solution for very large datasets, as it involves sorting the entire list and iterating through it sequentially.



Anonymous

Provide assert statements for the function. Include core and edge cases.

ChatGPT

Sure, let's add some `assert` statements to validate the function for core and edge cases:

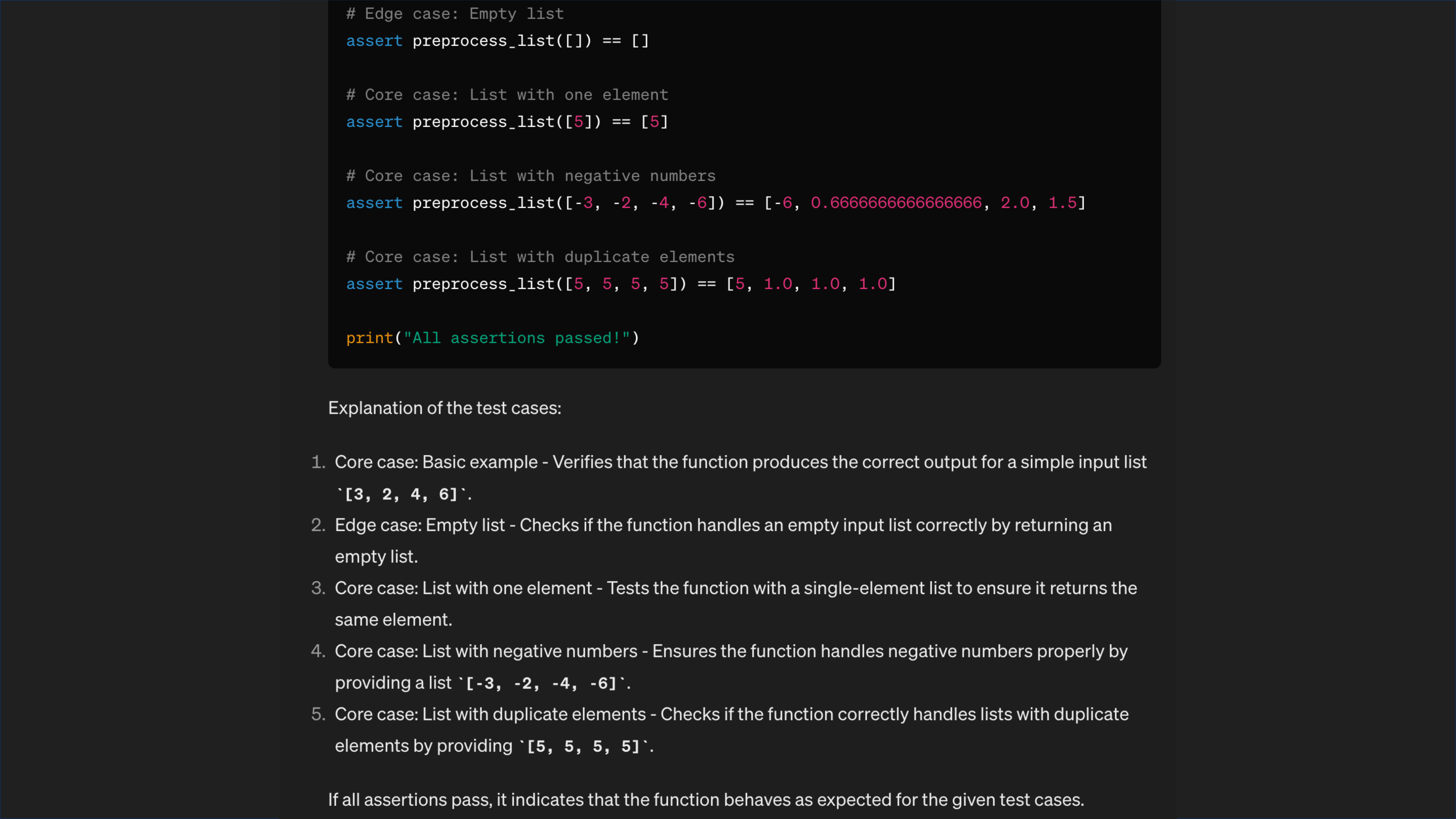
```python
def preprocess_list(numbers):
    # Sort the list in ascending order
    sorted_numbers = sorted(numbers)

    # Initialize a list to store the results of division
    result = []

    # Iterate through the sorted list
    for i in range(len(sorted_numbers)):
        # If it's the first number, append it to the result list
        if i == 0:
            result.append(sorted_numbers[i])
        else:
            # Divide the current number by the previous number and append the result
            result.append(sorted_numbers[i] / sorted_numbers[i-1])

    # Return the preprocessed list
    return result

# Test cases
# Core case: Basic example
assert preprocess_list([3, 2, 4, 6]) == [2, 1.5, 1.3333333333333333, 1.5]
```



```python
# Edge case: Empty list
assert preprocess_list([]) == []

# Core case: List with one element
assert preprocess_list([5]) == [5]

# Core case: List with negative numbers
assert preprocess_list([-3, -2, -4, -6]) == [-6, 0.6666666666666666, 2.0, 1.5]

# Core case: List with duplicate elements
assert preprocess_list([5, 5, 5, 5]) == [5, 1.0, 1.0, 1.0]

print("All assertions passed!")
```

Explanation of the test cases:

1. Core case: Basic example - Verifies that the function produces the correct output for a simple input list `` `[3, 2, 4, 6]` ``.
2. Edge case: Empty list - Checks if the function handles an empty input list correctly by returning an empty list.
3. Core case: List with one element - Tests the function with a single-element list to ensure it returns the same element.
4. Core case: List with negative numbers - Ensures the function handles negative numbers properly by providing a list `` `[-3, -2, -4, -6]` ``.
5. Core case: List with duplicate elements - Checks if the function correctly handles lists with duplicate elements by providing `` `[5, 5, 5, 5]` ``.

If all assertions pass, it indicates that the function behaves as expected for the given test cases.

Current Popular LLMs

ChatGPT
(OpenAI)

Copilot
(Microsoft)

Gemini
(Google)

Claude
(Anthropic)

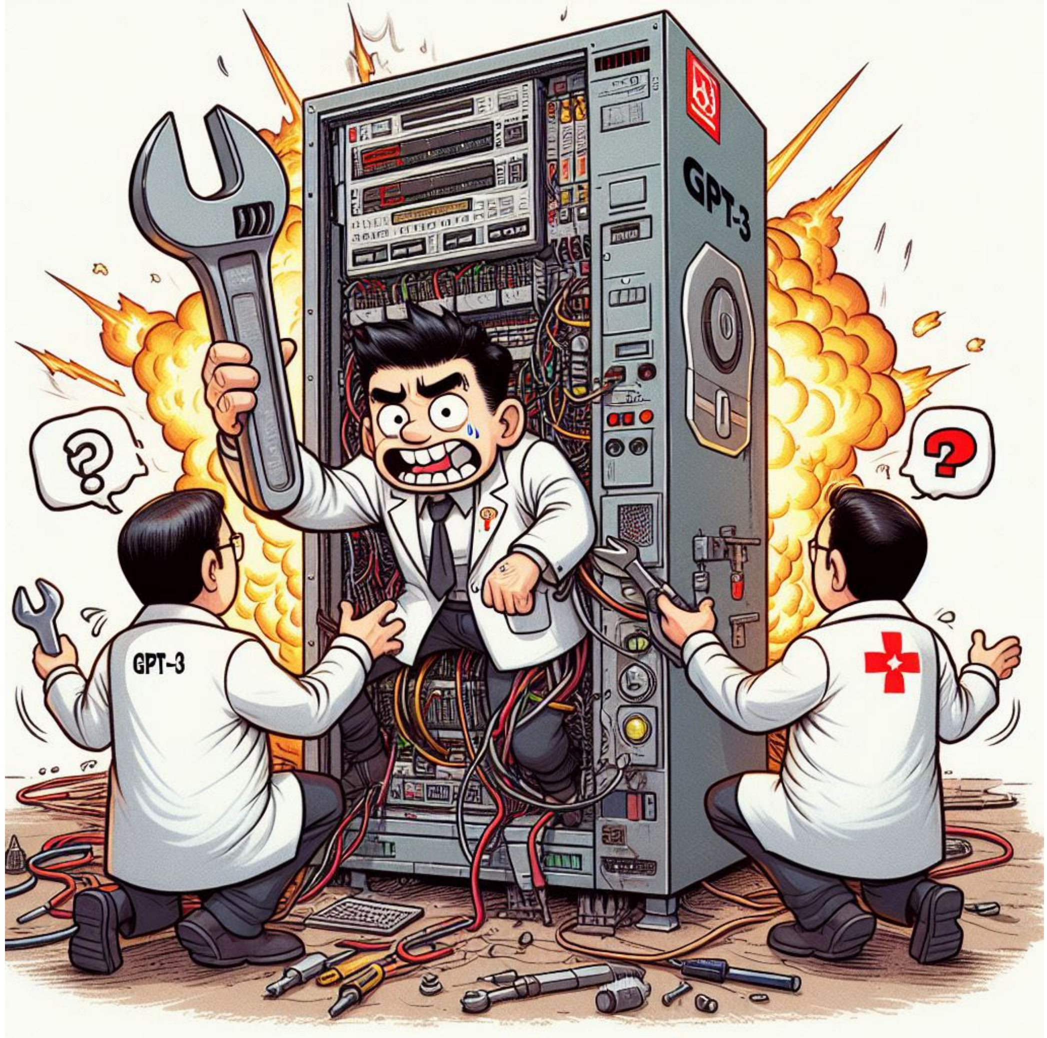
Hugging Face

Repository for MANY open-source GAI tools

Perils / Security Reminders

Perils of Coding with an LLM

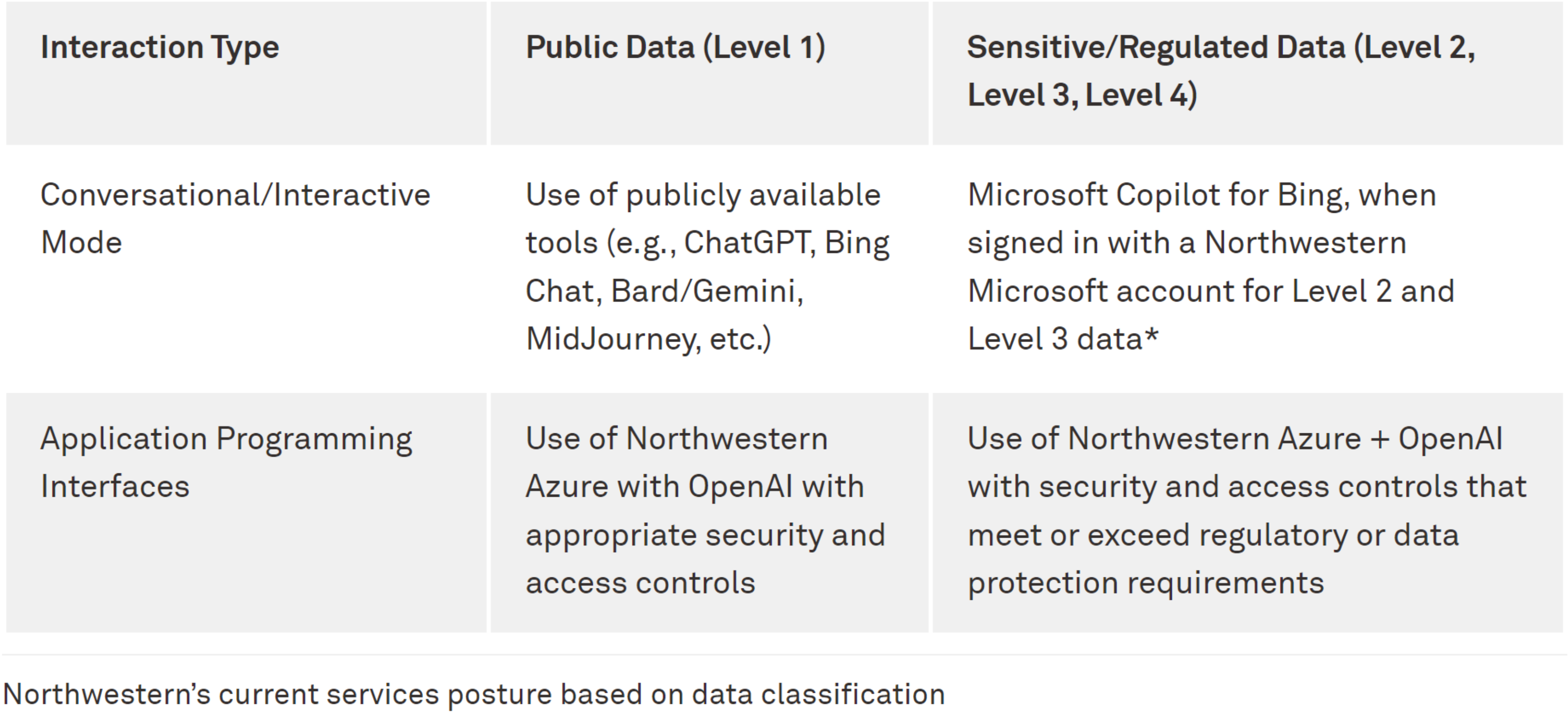
- **Bias**
  - Imai et al. (2022) shows gender bias increases (even for experts) after using an LLM to assist in an analysis task.
  - Don't blindly trust LLMs.
- **Intellectual Property**
  - If you use code produced from an LLM verbatim, do you own it?
  - If you share your code with an LLM, others may see/use it.
- **(Lack of) Privacy**
  - Do not share any data with an LLM that you would not also share with the public on the internet.
- **Cybersecurity**
  - Do not use code verbatim from an LLM without thorough testing and understanding.
- **Errors**
  - An LLM may suggest using erroneous attributes/ functions/ methods/libraries or other bugs. Always test the code!



^ Response from Copilot for "Please generate an image showing the perils of using a large language model."

Northwestern Guidance

https://www.it.northwestern.edu/about/policies/guidance-on-the-use-of-generative-ai.html



Interaction Type	**Public Data (Level 1)**	**Sensitive/Regulated Data (Level 2, Level 3, Level 4)**
Conversational/Interactive Mode	Use of publicly available tools (e.g., ChatGPT, Bing Chat, Bard/Gemini, MidJourney, etc.)	Microsoft Copilot for Bing, when signed in with a Northwestern Microsoft account for Level 2 and Level 3 data*
Application Programming Interfaces	Use of Northwestern Azure with OpenAI with appropriate security and access controls	Use of Northwestern Azure + OpenAI with security and access controls that meet or exceed regulatory or data protection requirements

Northwestern's current services posture based on data classification

Northwestern IT suggests using **Microsoft Copilot for Bing.**

If you have questions about your data security when using GAI, contact security@northwestern.edu

A Note on
“Vibe Coding”

How to review LLM-generated code if you don’t know the language?

- **Vibe Coding** is using an LLM to generate code based on (natural language) prompts, rather than writing code yourself.
- Though this makes coding (and different coding languages) more accessible to people w/o strong coding backgrounds, I have concerns:
  - All the perils from the previous section +
  - How to identify bugs / security concerns / etc. if you don’t know the code language?

How to review LLM-generated code if you don’t know the language?

- My suggestions:
  - ***Ask another LLM*** to verify the work of the first LLM and/or ask the same LLM in a different conversation.
  - Ask the LLM to ***describe each step in detail***. Be sure that you understand the logic of each step, even if you don’t understand all the code.
  - Include many ***print statements***, log outputs, and other tests in your code that may catch errors.
  - ***Read the documentation*** for libraries you are using in your code to be sure they are implemented correctly.
  - Use the LLM to ***teach you to code***, rather than producing production code for you.
  - ***Submit consult requests to our team!***

My final 2 cents:

If you want to do something with code, I strongly urge you to

learn the code language yourself

(at least enough to write the code you need for the task at hand). LLMs can help you learn, but don't rely on them to produce all your code.

Exercises

Your turn to work with LLMs (ChatGPT / Gemini / Copilot / Claude / etc.)

Exercise 1 - CPS Progress Report Analysis

- Use an LLM to generate code for a typical data exploration task: ***searching for correlations in a large data set.***
- We will use data from the Chicago Data Portal on Chicago Public Schools (CPS). The data is available on our GitHub repo **here**, in 2 files:

1. `‘Chicago_Public_Schools_-_School_Profile_Information_SY2324_20240924.csv’`
   demographic information

2. `‘Chicago_Public_Schools_-_School_Progress_Reports_SY2324_20240924.csv’`
   progress report results

Exercise 1 - CPS Progress Report Analysis

- Your task is to ask any LLM to generate code that will allow you to search for correlations in this data set (including potentially between demographics and progress). Then select one particularly strong and interesting correlation to share with us.
- Remember the suggestions for prompt engineering from earlier and see the **cheat sheet here**.

Directives

- You can use any programming language and any LLM.
- If you don’t know how to download the data – ask your LLM!
- Try solving this only with prompt engineering and not by coding the solution yourself.
- It is completely OK if you do not make it to the end of this exercise!

Exercise 1 - CPS Progress Report Analysis

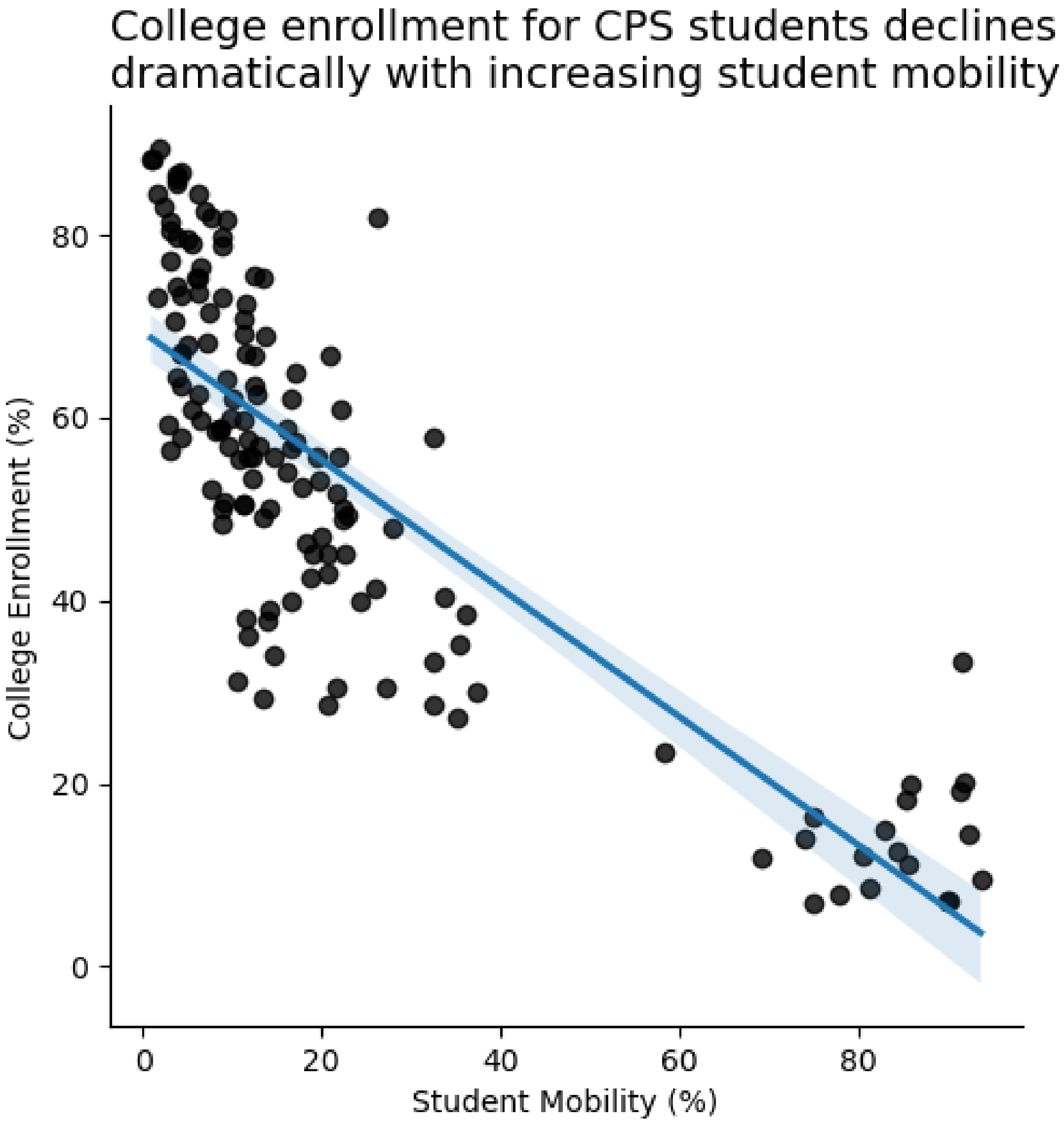
Hint: Break this task down into multiple steps, and ask the LLM for 1 step at a time:

1. Merge the 2 dataframes using the “School_ID” column.
2. (Optional) Create new columns in the merged dataframe that show the percentage of students in various categories, rather than the raw student counts (columns beginning with “Student_Count”).
3. Create a correlation matrix for all the columns that have numerical values.
4. Print out any strong correlations that seem interesting (e.g., that aren’t from similar metrics).
5. Check at least one of these correlations by creating a scatter plot that also shows a linear regression line.
6. Share the result with us!

Directives

- You can use any programming language and any LLM.
- If you don’t know how to download the data – ask your LLM!
- Try solving this only with prompt engineering and not by coding the solution yourself.
- It is completely OK if you do not make it to the end of this exercise!

Exercise 1 - CPS Progress Report Analysis



Exercise 2 – Refine your code

Use a different conversation, or a different LLM, to evaluate the code that you just generated. Here are some ideas to try:

1. Tell the LLM the purpose of the code, provide the code, and ask the LLM if it would recommend any improvements.
2. Ask the LLM to generate comments (or other documentation) for your code.
3. Ask the LLM to generate unit tests for your code.
4. Ask the LLM to suggest extensions to further your research on this data

Questions?

Thank you!