SEMROS GPT Prompts and Evolution

**About the prompts:**

Very early into the project, our team designed a web scraping function that uses the headless browser puppeteer to scrape the web and gather data related to a particular item. That data would be output in a very non-deterministic format. We needed some way to parse through this data, and although we considered not utilizing LLM at all, we decided the data was too diverse for handwritten algorithms. Therefore, we set up a system that simply asks ChatGPT to parse the data for us. This original prompt was quite basic:

<data from web scraping function here…>

Based on the text above answer the questions. Follow the format provided strictly. If you cannot complete the task respond with\\"impossible\" otherwise only provide the data. Prioritize higher precision answers in the text, ensure it is dimensions not $. Q: Provide the item's dimensions in the specified format\n";

As we tested this prompt, we realized that it would be quite difficult to get information relating to more subjective properties such as “fragile” and “stackable”. Matthew Periut decided to design a new system that allows the LLM to take control of our web scraping function. The resulting function is very dynamic and capable of finding information about dimensions, weight, fragility, stack-ability, and likely any number of other item data. This is the final system that is used in the SEMROS Product.

*Note these prompts were tailored to ChatGPT 4.0 Turbo specifically; however, the prompts may perform similarly with another LLM such as Claude or Google Gemini.*

**Current (Primary) Prompt**

The prompt below is the current data collection prompt and can be found on the continuous\_gpt.js file. It was designed with flexibility in mind, which means the prompt is dynamically adjusted based on the data that we are looking for. Information in blue may change depending on the parameters passed to its parent function.

You will be given a description of a physical item, and a list of properties to fill out. You can only reply with two things the first is google("question", "keywords"). If you respond this way, my function will search the internet using the question that you provide, evaluate the first 10 webpages, and return any text within 100 characters of the keywords. Use | to separate keywords. Example usage: google("how tall is mt everest?", "height|feet") If the results is [] it is likely that the keywords did not work. Please include item description in the question variable. and manufacturer part number as a keyword. the start of individual website data are marked by\"data\", validate the data by checking if the manufacturer part number is found on the data from that website. If you aren't confident in the data, adjust the keywords and try again. You have up to 4 google searches, so use them. When you have found the answer, you may use the second response: (variable as given): (answer, e.g. "true", "false", "number"). However, try not to answer what the variable is until you find it. If you can't find enough data, search again. Follow these guidelines strictly. On the final try you will be informed that you can no longer google search and must reply.

Item description is <item\_desc> Manufacturer part number: <manufacturer\_part\_num>

Property 1: <property> , Property Type: <type> additional info is: <Additional\_info >

Property 2: <property> , Property Type: <type> additional info is: <Additional\_info >

Properties are appended to the prompt depending on the use case.

**Secondary Guidance Prompts**

These prompts are added to the message “stack” in order to guide the LLM through different control flows.

**Correct Scraping Function call:**

Results for "${question}" with keyword "${keywords}": <output>

**Incorrect Scraping Function call:**

Sorry, your response did not match the expected format. Please reply with 'google(question, keyword)’

**Out of scraping Attempts:**

You must answer now, googling is disallowed. respond (variable):(answer)

**Correct Output Call:** Processed your Response for <property> Good Job, please answer the next property, <property>.

**Additional Information Parameters**

This data