AI Prompts Developed For and By Researchers

Notes for Use: *We tested research prompts on Large Language Models (LLMs) to leverage LLMs in our research work. These prompts were tested using GPT-4 and/or Claude 2. Prompts follow the* ***bold blue headings*** *and can be copied and pasted into an LLM; the orange text indicates wording that should be customized.*

Parse a String Variable for Analysis

The variable in the attached dataset combines the calendar date and the time of day. Parse the date out into a new analytic variable. Then graph the new variable so there is a visual display of when animal bites occurred.

Teach an LLM to Write in the Style and Format You Need

Prompt 1: Read the attached PDF, paying close attention to writing voice and style, level of concision and clarity, reading level, and other features. Remember all details of this style as "scistyle". When you are done reading and feel you have captured "scistyle" reply "done" and nothing else. If you need more information to capture "scistyle", reply with a question that will help me provide what you need.

Prompt 2: After "done", paste or upload the text you want edited and prompt with "Rewrite this prose using "scistyle".

Merge Datasets and Download the Merged File

Attached are two datasets, one contains hospital admission records and the other contains COVID cases, with unique identifiers. I am interested in knowing how many COVID cases ended up hospitalized or not; suggest which join to use. Then merge the two datasets by "Case ID" and display a count of unmatched values, if any. Name the dataset that contains IDs from both files "mergedData" and present it for download.

Generate Descriptive Statistics and Format Them Into a Table

Using the attached dataset, create a table of descriptive statistics of COVID cases stratified into two columns by "Hospitalized" using the following guidance:

1. Include total counts for each column.
2. Describe the statistics of each variable as rows. Display the mean and standard deviation for continuous variables; include the count and percentage for categorical variables; include all unique values of categorical variables as their own row.
3. Using an asterisk or other symbol, footnote missing values and then place the footnote as a count of the number of missing values at the bottom of the table where applicable.
4. Present counts as whole numbers and everything else rounded to two decimal places.
5. Give the table and the row headers in the table an informative title.
6. Output a detailed summary interpretation of this table.

DATA-DRIVEN LLC

Create a Visual Display of Data

Visualize the pattern of crime by day of week and month of the year in the attached. Visualize this as a heat map. Give the visualization an informative title. Present it for download.

Subset Data Using Specific Criteria

For the dataset attached, I am interested in a new dataset with only those observations for which hospitalized status is not missing and where the test date is after Feb 1, 2020, and no later than October 1, 2023. Name that new dataset and present it for download.

Create a Codebook by Uploading a Survey Instrument to Claude

Attached is a survey data instrument. Create a formatted, user-friendly codebook for this instrument using the following guidance:

1. Create short and intuitive variable names.
2. Code dichotomous variables of 0/1 where 0=no, and 1=yes.
3. Where there is a note about "Skip Logic", note which question might be skipped and create a skip logic code for that question that is explicitly marked "skipped due to logic".
4. Combine "missing" / "prefer not to answer" into one code you define.
5. Define a separate code for "unknown" and another code for "not seen due to skip logic".
6. Be sure to include codes for all questions in the attached instrument.
7. Do not abbreviate or use string values for response codes.

Caveats

- **LLMs create new content**: Generative AI has an element of randomness, meaning that repeating the same prompt will yield similar but not identical results.
- **The only constant is change:** Since models change rapidly these prompts may need modification over time.
- **Hallucination:** A helpful (but not failsafe) check against hallucination is to challenge the LLM, and insist it show you how it got to its output.
- **Verify, don't trust:** Ask the LLM to gauge its capabilities in tasks that require accuracy. Check all output and more closely check the output from high-difficulty tasks or skip these tasks until models improve.
- **Codebooks:** Include skip pattern documentation in codebook prompts. Codebooks were a homerun for Claude, but not GPT-4. We recommend Claude for codebooks.
- **QA hack:** Ask the LLM to offer counts of results of analytic work, e.g., in a merge, ask how many unmatched values were found. Or for descriptive stats, ask how many instances it finds of the quantity listed as the mode. Do this in two LLMs and compare results.
- **Stats & data viz:** We detected substantial hallucination with Claude's descriptive statistics and limited capability with data visualization. Consider using GPT-4 with Advanced Data Analysis for all data analysis until other models improve.

DATA-DRIVEN LLC

For detailed LLM output and our most recent updates, email: leslie@wearedatadriven.com