DAT565/DIT407
Introduction to Data Science and AI

Assignment 7

Deadline: 2024-10-24 08:00

In this assignment, you will investigate using a large language model to understand its strenghts and weaknesses. The aim of this assignment is to reflect and document your experience. Hence, there are no black and white, right or "wrong" answers to each question, so long as they clearly explain what you did and your analysis is not too superficial. You will pass if you complete each task and write a clear, thorough analysis of your findings. This can be a few short sentences or a paragraph, but they must contain insightful and well-written reflections on the results. The number of sentences or words stated is a guideline; if you want to say a little bit more that, that is OK. You should write your reflections yourself, not using the chatbot (and after completing the assignment, you may also realize why this is a good idea). For some background reading, you may want to also read the article about ChatGPT failures [1].

Task 1: Choose a chatbot

First of all, choose a publicly available chatbot for your experiment. You may choose between these two systems that are free:

- Llama 3: `https://www.llama.com/`
- ChatGPT 3.5: `https://chat.openai.com/`

Note that ChatGPT requires a login. If you are not comfortable with this, choose Llama 3. If you use Llama 3, make sure to specify exactly which version you used in your report. If you want, you may also try the same questions with both systems and compare the results.

You are advised not to share personal details in the chatbot. Only type in text that you are prepared to allow being used in further training of these systems. To answer Task 1, clearly state which chatbot(s) you are using for this assignment and how you accessed it.

Task 2: Find a question to which you get a factually incorrect answer

(a) Experiment with your chosen chatbot and ask it questions in English. Try to invent a question for which the chatbot gives a factually incorrect answer. In your answer, state the question and what the chatbot replied.

(b) Explain what part of the answer was factually incorrect.

(c) Did you find it hard or easy to come up with such a question? Why/why not? [Write 2-5 sentences]

(d) What could be reasons for the chatbot producing a factually incorrect answer? [Write 2-5 sentences]

Task 3: Exploring bias and stereotypes

Ask the chatbot to describe a person. For example, you may ask it to *"Describe a character who is a CEO of a large tech company"* (but it is more fun if you come up with something yourself).

(a) State your prompt to the chatbot, and its response. You may include one or two different prompts and answers.

(b) Does the description(s) contain anything that can be described as biased or stereotypical? Or something on the contrary that was novel and unexpected? [write 100 – 300 words]

(c) Does the chatbot do a good or bad job avoiding stereotypes and bias, in your opinion? Why do you think that is? [Write 2-5 sentences]

(d) What happens to the answer if you add to the end of your prompt a sentence *"Be creative and make it less stereotypical"* and/or *"Make it very stereotypical"*. Does this change the answer? How?

Returning your report

Write a report, typeset in LaTeX, that answers *all* questions above. Include all your Python code in your report as an appendix, preferably using the `listings` package. Your report should be legible even without having a look at your code.

If you refer to outside sources, remember to add an appropriate literature reference (including websites) in references by `\cite`ing the references. It is recommended that you use the package `biblatex` to manage citations.

Place your figures in numbered `figure` environments, with descriptive captions and `\ref` to the figures in your discussion. Likewise, place your tables in numbered `table` environments with desciprive captions and `\ref` to the tables in your discussion.

After grading, you will be given another attempt to revise your report according to TA comments if it is not considered acceptable.

The deadline is *hard*. Late submissions will not be read at all and are considered failed. This means you will not get any feedback for the first round and the submission is considered a revision; there will be no third attempt, so if a late submission is failed, you will need to participate in a later iteration of the course for a re-attempt.

References

[1] Ali Borji. "A Categorical Archive of ChatGPT Failures". In: *CoRR* abs/2302.03494 (2023). DOI: `10.48550/ARXIV.2302.03494`. arXiv: `2302.03494`. URL: `https://doi.org/10.48550/arXiv.2302.03494`.