Instructions: Evaluation of GPT 4, LAMA 2, and CLAUDE 2 models

In the context of this project, you have to assess three natural language processing tools: GPT 4, LLAMA 2, and CLAUDE 2. Your task is to evaluate the performance of these tools based on the provided tests.
You can use a free VPN to access Claude 2 (https://chrome.google.com/webstore/detail/urban-vpn-proxy/eppiocemhmnlbhjplcgkofciiegomcon).

In this Excel file you find your topic and the model assigned to you:
https://docs.google.com/spreadsheets/d/1_m01IU7jl-dbLAOcTkLJ7qjfgQpF8WfPHkISll8MTQE/edit?usp=sharing

Since almost no one filled out the file specifying which model they would use, we randomly assigned the model. However, groups with the same topic can decide to change the model. for example, if Groups 1 and 2 both work on questionnaire X, you can choose who will use Llama 2 or who will use Claude 2. The important thing is that you mark it on the excel file in order to keep track and avoid having 2 or more groups working on the same topic with the same model.

Based on the name of the topic, you will find the material here:
https://drive.google.com/drive/folders/1HPBMKWYVyXbEI5gspDEIUa7Itw5C7jUf?usp=sharing

With the exception of the groups with the *lie detection* theme which will be explained later in this document and the group with the *hallucination* theme in LLm who have already received the material from the professor, those who will work on the theme: *prefrontal function, analogical reasoning and problem solving*, will have to read the papers that they will find in their topic folder. You will then have to apply the same procedure described by the authors and test the model assigned to you (e.g. Claude 2). Ultimately you will have to compare the model's performance with human's performance. In some cases the authors also measure and evaluate human reaction times. Ignore this information, you have to evaluate models only on the accuracy of the response.
In some cases in the folder you will also find an additional paper in which the authors have already applied that test to another model (different than yours). Use that paper as a starting point to build your report or to have extra usefull information.

Output
Each group must prepare a report on the results obtained. In the following link, we provide an example of how you can structure the report.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4377371

Additionally, you should provide a well-organized tabular file containing the item test, the raw answers provided by the tool, and the scores/interpretations you assigned to the responses. You must submit both the report and the results file by the oral examination date (send an email to giuseppe.sartori@unipd.it and giulia.melis@phd.unipd.it with all the materials). During this examination, you will individually present your work.

In the last four lessons of December (13, 14, 20 and 21), each group will present the status of your project. You are not required to present complete results, but rather the current progress, to receive feedback from your colleagues and guidance from Prof. Sartori.
For both presentations, we recommend using slides.

Lie Detection Tasks
(ONLY FOR "LIE DETECTION" GROUPS)

Lie detection tasks are divided into two groups: fine-tuning a LLM or prompt-engegneering a LLM.

For fine-tuning a LLM and understanding what actually is a Scenario 1 or Scenario 3, you have to carefully study this paper and check this git-hub repository:

https://www.researchsquare.com/article/rs-3126100/v1

https://github.com/robecoder/VerbalLieDetectionWithLLM

You may struggle in fine-tuning a LLM using Colab and you may require to use Colab Pro+. If this is the case, you can decide to:

a) access to Colab Pro +
b) choose a smaller model for the fine-tuning
c) choose a smaller portion of the dataset

Please, describe and motivate appropriately what was the process you followed and why you decided to proceed in this way. It is also recommended to coordinate with colleagues from other groups that works on the fine-tuning of the same LLM but on a different dataset (e.g. group a will fine-tune Llama2 on Opinion dataset, group b will fine-tune Llama2 on Memory dataset).

For the prompt-engegneering, you are asked to test different prompting strategies, from the simplest to the most sophisticated ones.

I would recommend starting with zero-shot learning, then few-shot learning, then chain-of-thought (if applicable), self-consistency (with decision based on voting), tree-of-thought.

It is recommended to build a code to test those procedures on multiple examples. You can also decide to test the prompting strategies on a subset of the whole dataset (e.g., 100 examples). Please, decide together with your colleagues a reasonable number of examples to be tested and try to be consistent among the test set selected.

Example: If group a, b, c, are testing prompting strategies in Scenario 1 with the Opinion dataset using three different LLMs (e.g. group a using Llama 2, group b using Claude 2, group c using GPT4), please create one common test set of e.g. 100 examples, so that at the end the results of the performances on the different models and different techniques are fairly comparable.