**CS521 Project -Investigating Bias and Memorization in Language Models: An Analysis Across Diverse Question-Answer Datasets**

**Description:**

This project explores bias and memorization phenomena in language models (LMs) using various question-answer datasets. We analyze how LMs respond to different types of questions, considering factors such as wording variations, demographic references, question structures and jumbling options. The analysis aims to provide insights into the behavior of LMs regarding bias, fairness, and memorization across diverse datasets.

**Novelty:** The novelty in this project lies in its comprehensive analysis of bias and memorization in language models (LMs), particularly ChatGPT and Claude, using diverse question-answer datasets. Here's how the project introduces new techniques or applications and conducts novel comparisons:  
  
1. **Integration of Language Models with OpenAI API:**

* The project leverages the OpenAI API to interact with state-of-the-art language models, including ChatGPT and Claude. This integration allows for the evaluation of these models' performance in generating responses to various survey questions.

2. **Analysis Across Diverse Datasets:**

* The project analyzes bias and memorization in language models across multiple datasets, each presenting unique challenges and characteristics.
* Datasets such as allow\_forbidden, demographic\_rephrase, and jumbled\_options offer different perspectives and complexities, enabling a thorough examination of LM behavior.

3. **Bias Detection and Fairness Measures:**

* Through meticulous data processing and analysis, the project identifies and quantifies bias in the language models' responses.
* Fairness measures, such as comparing responses to original and scrambled questions, provide insights into the presence and extent of bias in LM outputs.

4. **Memorization Analysis:**

* The project goes beyond bias detection to investigate memorization in language models.
* Novelty scores and overlap analyses offer unique perspectives on the diversity and consistency of LM responses, shedding light on their memorization capabilities.

5. **Comparative Analysis Between Models:**

* By comparing the performance of different language models (e.g., ChatGPT vs. Claude), the project offers insights into their strengths, weaknesses, and overall suitability for various applications.
* Novel comparisons highlight differences in bias detection, memorization tendencies, and overall response quality between the evaluated models.

**Source code Description:**

1. OpenAI API Integration:

* The code integrates with the OpenAI API to interact with language models (LMs) such as ChatGPT and Claude.
* It utilizes the OpenAI Python package to facilitate communication with the API.

1. Prompt Generation and Response Retrieval:

* The code generates prompts for the language models based on provided format instructions and survey questions.
* It retrieves responses from the language models using the generated prompts.

1. Data Processing and Analysis:

* The code performs data processing tasks such as cleaning and standardizing survey responses.
* It analyzes the data to measure bias and memorization in the language models' outputs.

1. Visualization and Reporting:

* The code generates visualizations (e.g., bar charts, pie and stacked bar charts) to present the analysis results.
* It computes metrics such as fairness measures, novelty scores, and overlap analyses to quantify bias and memorization.

1. Output File Generation:

* The code generates output files (e.g., CSV files) containing processed data and analysis results.
* It saves the output files for further examination and reporting.

**Data Collection and Preprocessing**

**We manually built the input datasets by surveying people.**

1. allow\_forbidden.csv: This dataset contains pairs of questions, with one question presenting a statement in an affirmative stance, while the corresponding "scrambled" question presents a similar statement but from a negative or opposing perspective.
2. demographic\_rephrase.csv: The demographic\_rephrase dataset consists of survey questions related to various demographic factors, such as location, socioeconomic status, and personal opinions or concerns. Each question is paired with a corresponding scrambled version, altering specific details while maintaining the overall structure and context.
3. jumbled\_options.csv: This dataset contains questions where the options for response are presented in a different order compared to the original question.

**Input Data Format:**

1. allow\_forbidden.csv: CSV file with pairs of questions and scrambled questions.
2. demographic\_rephrase.csv: CSV file with survey questions related to demographic factors and their scrambled versions.
3. jumbled\_options.csv: CSV file containing questions with jumbled response options.

**Source Code Implementation:**

**Bias Analysis**

* GPT LLM Bias Analysis: We utilized the GPT LM (Language Model) to analyze bias in responses. The LM generates responses to survey questions, and we compare the similarity between responses to assess bias.
* Fairness Measure Calculation: We calculate the percentage of responses that match exactly, indicating similarity between responses.
* Claude Bias Analysis and Fairness Measure Calculation: Similar to the GPT LLM Bias Analysis, we conducted bias analysis using the Claude LM to compare responses and calculate the fairness measure.

**Memorization Analysis:**

* GPT LLM Memorization Analysis: We examine memorization behavior in the GPT LM by analyzing the consistency of responses across different outputs.
* Data Merging: Multiple output files are merged based on question and scrambled question pairs.
* Memorization Identification: We identify memorization by checking if the three answers are the same for each question.
* Novelty Score Calculation: We calculate the novelty score to assess the uniqueness of responses.
* Overlap Analysis: We analyze the consistency or redundancy in responses using the overlap analysis score.
* Claude Memorization Analysis: Similarly, we conduct memorization analysis using the Claude LM by merging output files and calculating novelty score and overlap analysis.

**Stereotype Analysis:**

* Examining if there are any stereotype words present in the question and if they are present, finding out if the answer was same for the actual question and scrambled question for those type of questions
* Conducted the above anaylsis for both chatGPT and Claude model for the demographic dataset

**Instructions for Running the Source Code:**

1. Install the required Python packages using pip:

* pip install openai
* pip install anthropic
* pip install pandas
* pip install plotly

1. Replace the file paths in the code with the paths to your dataset files.
2. Ensure you have valid API keys for accessing the OpenAI API (for ChatGPT and Claude).
3. Execute the provided Python script to run the bias and memorization analyses.

**Output and Results files**

* Bias Analysis Output: The output of the bias analysis includes fairness measures calculated for GPT LLM and Claude.
* Memorization Analysis Output: Output files contain analysis results, including novelty scores and overlap analysis scores, for both GPT LLM and Claude.
* Stereotype/Demographic Analysis Output: Output files contian the stereotype analysis output along with visualizations showing number of matched answers for the stereotype questions
* The results files are allow\_forbid, jumbled\_options which are generated by feeding them to the LMs.

1. allow\_forbid Folder:

* output\_allow\_forbid\_\*.csv: Output files containing responses generated by both the GPT LLM and Claude LLM for the allow\_forbid dataset.

1. jumbled\_options Folder:

* output\_jumbled\_options\_\*.csv: Output files containing responses generated by both the GPT LLM and Claude LLM for the jumbled\_options dataset.

1. demographic Folder:

* output\_demographic\_rephrase\_\*.csv: Output files containing responses generated by both the GPT LLM and Claude LLM for the jumbled\_options dataset.

**Extent of Contributions:**

1. **Objective Fulfillment:**

* The project effectively addresses the goals outlined in the project proposal, including the analysis of bias and memorization in language models across diverse datasets.
* It systematically evaluates the performance of ChatGPT and Claude, comparing their responses to different types of questions and scenarios.

1. **Comprehensive Analysis:**

* The project conducts a thorough examination of bias and memorization in language models by utilizing multiple datasets and employing various analytical techniques.
* It covers different aspects of LM behavior, such as response diversity, consistency, and memorization tendencies, providing a comprehensive understanding of their performance.

1. **Integration of OpenAI API:**

* By successfully integrating with the OpenAI API, the project demonstrates the ability to access and utilize state-of-the-art language models for research purposes.
* Despite limitations with other LLMs like Google Gemini and PaLM due to unavailable APIs, the project adapts and focuses on ChatGPT and Claude, ensuring the completion of the proposed objectives within the available resources.

1. **Limitations Acknowledgement:**

* The project acknowledges limitations, such as the unavailability of APIs for certain language models, and adapts its scope accordingly to focus on available resources.
* Despite these constraints, the project effectively utilizes the available APIs to conduct meaningful research and draw insightful conclusions.

**We were able to get the APIs which were publicly available which were ChatGPT and Claude by purchasing them, but we tried for other LLMs like Google Gemini, PaLM but their APIs are not publicly available, so we linited our research with ChatGPT and Claude**