Model_Testing

1. Model_Response_Generation.ipynb: Contains scripts to generate prompts for the BBQ benchmark
2. Evaluate_Outputs.ipynb: Contains scripts to score the outputs of the model based on the prompts
3. Prompts: Contains jsonl files that store the prompts used to run the tests
4. Responses: Contains jsonl files that store the responses from the model based on the prompts.

Red_Teaming

Each use case sheet contains subfolders for different aspects of the use case (Low Context, Pros & Cons, Counterfactuals, Role Playing, Role Playing x Counterfactuals, Role Playing x Pros and Cons), each divided by attributes to target different groups such as Gender and Race, etc.

1. Financial_Advisement Use Case: Use case scenario focusing on fairness in finance context, such as home loans, tax planning, debt management, etc.
2. Payroll_Chatbot Use Case: Use case scenario focusing on fairness in chatbot interactions regarding payroll issues such as paid leaves, overtime, raises, etc.
3. Purchase_Recommendations Use Case: Use case scenario examining fairness in AI-driven recommendations for any product or service
4. Odds: Calculates the odds ratio which provides insight into how frequently the model generates a red response compared to generating either amber or green responses
5. Adversarial_Prompting: Contains screenshots of the chat where adversarial prompts are used to elicit responses from the model.

General Structure for Each Use Case:

- Low Context: Files analyzing minimal context scenarios
- Pros & Cons: Analysis of advantages and disadvantages for different groups
- Counterfactuals: Examination of outcomes if different choices had been made
- Role Playing: Scenario-based roleplaying to understand different perspectives