Readme

**Overview**

This document details the contents and instructions for the client submission package focused on model testing and red-teaming analysis for fairness in AI applications.

**Contents**

**- Model\_Testing**

* Generating\_responses\_from\_model.ipynb: Contains scripts to generate prompts for the BBQ benchmark.
* Evaluate\_Outputs.ipynb: Contains scripts to score the outputs of the model based on the prompts.
* Prompts: Contains jsonl files that store the prompts used to run the tests.
* 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.

* Financial\_Advisement Use Case: Use case scenario focusing on fairness in finance context, such as home loans, tax planning, debt management, etc.
* Payroll\_Chatbot Use Case: Use case scenario focusing on fairness in chatbot interactions regarding payroll issues such as paid leaves, overtime, raises, etc.
* Purchase\_Recommendations Use Case: Use case scenario examining fairness in AI-driven recommendations for any product or service.
* 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.
* 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.