INTERIM REPORT

This project explores whether large language models, like ChatGPT, Claude, and Gemini show bias when asked to analyze the same data in different ways. The goal is to see if changing how a question is framed for example, calling a player "underperforming" versus "developing" affects how the model interprets the same basketball statistics.

I used anonymized player data from the Syracuse Men's Basketball team as the dataset. I designed five small experiments that test for different kinds of bias, such as framing effects, demographic influence, and confirmation bias. For each experiment, I created two or more prompts that are almost identical except for one small difference in wording or context. These prompts are then given to different language models, and their responses are collected and analyzed.

I built a few simple Python scripts to help with this process. One script creates all the prompts automatically, another runs the experiments and saves the model responses, and others analyze the results and check them against the real data. This setup ensures that every step is consistent and easy to repeat later.

Early testing suggests that how a question is asked can change the tone and direction of a model's response, even when the data itself hasn't changed. For instance, positive wording tends to produce more encouraging answers, while negative wording leads to more critical ones.

The next stage of the project will involve running more tests across multiple models and comparing results. I will also measure how often the models exaggerate, make assumptions, or overlook parts of the data. The final goal is to understand how subtle wording choices can lead to bias and to identify practical ways to make model-based analysis more balanced and reliable.