Research Task 5:

Descriptive Statistics and Large Language Models

Welcome

This period's task is to take a small public dataset (potentially from a past season of an SU sports team) and provide that dataset to a large language model like Co-Pilot, Claud, or ChatGPT and then challenge the LLM to correctly answer natural language questions about that data. I will be completing this task using the SU Women's Lacrosse data from the past season:

https://cuse.com/sports/2013/1/16/WLAX_0116134638

Every SU sport has a similar yet distinct set of data. Alternatively, use data from your favorite Cricket team or any other public dataset you like. Keep it on the smaller side so you don't blow through credits with a giant context.

I am sure you will have to do some novel prompt engineering to get some of your questions answered and answered correctly. I am assuming questions like "How many games did this team play?" will be answered correctly on the first go, but more probing questions like "Who was the most improved player this year?" will involve having to define a measure of quality. Please use your prior work to validate that the answers the LLMs are providing are actually correct. For those joining us this period, please see a classmate for an example set of scripts (or write your own) to generate basic statistics.

This assignment is more research oriented and let's take two reporting periods (at least) to work through it. Note, I am more interested in your experience here than if your can brute force the LLM into a correct answer. It is just as meaningful to me to know that an LLM failed at this task, what you tried and how it failed. As a reminder, some LLMs will execute python scripts and that might also be helpful.

I'll leave you to determine the natural language questions you ask of your data but I'll ultimately be attempting to get the LLM to provide answers to questions like "As a coach, if I wanted to win two more games this coming season, should I focus on offense or defense and If so, what is the one player I should work with to be a game changer and why?" There is a lot going on there that likely requires be to establish metrics that the LLM needs to make these kinds of judgement calls.

Any visualizations you can produce directly from the LLM would be great to see.

Again, this will be a longer more research intensive project, so we will work of two reporting periods at least to see what we can come up with. Two important reminders though. The first is you must report your activity twice a month (July 31st and August 15th) even though this specific task is longer than a period. The second is to be sure that when communicating with me you use my email (jrstrome@syr.edu) and not that of Dr. Stromer-Galley who as an admittedly very similar name and address.

Submission Instructions

- Create a **public GitHub repository** titled: `Task_05_Descriptive_Stats`
- Include support material such as scripts and prompts and a readme.md
- `As a reminder,` **do not include the dataset file** in your GitHub repo.
- Submit your repository link to jrstrome@syr.edu

Time Reporting Requirement:

It is critical that you report your research progress via the Qualtrics survey. This is the main way we are tracking OPT activity for when we must report to the government. Please ensure you complete these check-ins in a timely manner.

https://syracuseuniversity.qualtrics.com/jfe/form/SV_cDgnzM695AMx8d8