BREAK THROUGH TECH

Team Progress Summary

I. TEAM INFO

- AI Studio Host Company: Relativity
- Challenge Project Title: Evaluating Frontier Models for Performance and Representation
- Challenge Advisors: Mary Gibbs, Sean Underwood
- AI Studio TA / Tutor / Course Support: Preston Firewood
- Student Team Members:
  - Ana Garcia
  - Kennedy Martin
  - Jannatul Nayeem
  - Helen Song

II. CHALLENGE PROJECT ACCOMPLISHMENTS

Our project focuses primarily on research and exploratory analysis into detecting model bias and then mitigating it, rather than on building a model. Due to this, the majority of our time to this point has been on researching different bias detection methods and their corresponding mitigation techniques as well as researching different potential datasets to use and attempting to manipulate their data.

We have reviewed literature on measuring bias–different benchmarks and task evaluation (ex. Question answering, NER, sentiment analysis). We have also researched several different datasets, comparing them, and brainstorming different ways that we might be able to expand or change them in some way to highlight a new bias. Furthermore, we have worked on loading and manipulating data acquainted with the datasets in Google Colab and in the process have learned Hugging Face as well as LangTest.

The initial few weeks were centered around understanding how bias can manifest and be measured in LLMs, as well as determining different directions for our own project. Most recently, we've honed in on NER (Name Entity Recognition) in the legal space, and are aiming to restructure a GitHub dataset for our own evaluation with substituting different identities (ex. Changing pronouns, race categories, ethnic backgrounds, countries).

<u>Technical Artifacts</u>:

- **GitHub Repo:** Relativity GitHub Repository
- **Python Notebook(s)**: Google Colab (using Langtest demo)

III. CHALLENGES AND SOLUTIONS

Describe any significant challenges your team has faced and how you have addressed them. Share any important lessons learned that can help guide future steps in the project.

Challenge #1: Project definition and scope

Initially, we had a hard time understanding the specific expectations of how we would be investigating bias–was this building off of the initial papers that our CA's provided, or were we to come up with an entirely novel way to assess bias? What was the scope of the project–did it have to concern RAG and the legal space?

Our Solution for Challenge #1:

We met with our CA's to clarify these questions, and now understand that the task is very open ended and up to our group to decide what to focus on. Although the goal is to find a somewhat unique angle on how to investigate bias in LLM's, the main priority is to find any interesting insight in the context of bias assessment within the timeline of our project.

Challenge #2: Running code from literature locally

Furthermore, our team struggled with understanding how to load different datasets and then properly replicate results and experiment with them on our local machines.

Our Solution for Challenge #2: Our TA, Preston, gave us pointers on the types of helpful resources and how to use them–HuggingFace, how to use API keys, LangTest. We also met synchronously, sharing our screens to work through replicating the BBQ dataset together.

Challenge #3: Data modification/creation

We were unsure of how to best modify existing datasets to arrive at a dataset ready for model testing. We started by creating different sets of prompts based on indicators of both low and high economic status, but weren't getting differential responses from the LLM.

Our Solution for Challenge #3:

We met with our CA's and TA, and decided it would be easiest if we created or found a template which would change in a standardized way based on different conditions. We also discovered that the measurement technique being used wasn't the most efficient and were suggested some alternatives.

Lessons Learned:

- Constant communication and status updates about what is going well/what is difficult and what still needs to get accomplished is crucial.
- Working together synchronously and screen sharing can be a great way to get everyone on the same page when struggling with something technical.
- Taking time at the beginning of a meeting to check in on each other outside of AI Studio is critical in our mindset in the project and success.

IV. NEXT STEPS

Outline the next steps your team plans to take in the coming month. What are your goals and what tasks will you focus on? How will you divide up and/or collaborate on these tasks? Are there any areas in which you could use additional support from your AI Studio TA or Challenge Advisor? Be specific about the tasks and the support needed.

Goal's for October:

- **Select a dataset**
    - We will look into the Legal_NER dataset to see if it is suitable for bias evaluation. We intend to focus on evaluating the dataset through LangTest, an open-source python library for evaluating LLMs.
- **Define performance metrics**
    - Make a confusion matrix to get false-positives, false-negatives, etc.
    - Define how to measure bias and agree on performance metrics.
- **Construct a prompt that does NER**
    - Research best practices for creating prompts to ensure robustness.
    - Focusing on one entity type at a time (multiple entities at the same time can make performance metrics more challenging)

Task Division: All members

- **Research:** Preliminary research on how to classify bias and how other people are using the dataset.
- **Prompt Construction:** Create a few prompts and then convene as a team to review and discuss thoughts.
- **Documentation:** Continue documentation through meetings as well as maybe add a time during our team meeting to check over the schedule for upcoming documentation submissions.

Support Needed:

- To stay on track, we will be updating the CAs and TAs with our progress and findings to ensure that we answer any pending questions and get feedback through Slack/Discord as soon as possible.

V. INDIVIDUAL CONTRIBUTIONS

Each team member should answer the following questions in their own words. Do not write anything on behalf of your teammate(s). In order for your team to be given credit for this assignment, a response must be present below from each team member. Each team member should reflect on their contributions and learning experiences honestly.

1. How have you contributed to project work so far?
2. What will you be focused on going forward?
3. What is one new thing you've learned from working on this project (technical skill or teamwork/project management skill)?
4. What is one area in which you think you could improve?

Helen Song:

Contributions:

- Created and added to research document to keep track of papers and points for inspiration
- Created Notion for organizational infrastructure
- Came up with ideas for bias evaluation (resume screening, social media/news article scraping)

Focus Going Forward:

- Helping restructure NER dataset
- Help with using langtest to evaluate models
- Look into existing literature on topic and find any opportunities to investigate a new angle

New Learning: Learned about HuggingFace API, tools for bias evaluation like LangTest, learned how to use Google Colab

Area for Improvement:

- Do more technical data wrangling and exploration with models on Google Colab
- ^applying more of the research I've conducted into practical exploration and analysis

Ana Garcia

- **Contributions:** Assisted in documentation throughout each meeting, evaluated datasets through Google Colab in order to gain insight on which datasets we should/should not use
- **Focus Going Forward:** Look into how we can evaluate bias of our model and what type of changes we need to make in our dataset to get the most out of it.
- **New Learning:** Learned techniques on how to use datasets/models from Hugging Face and importing them to Google Colab to evaluate them.
- **Area for Improvement:** Improve my skills in importing datasets in LangTest in order to use it for bias evaluation.

Jannatul Nayeem:

- **Contributions:** Worked on research, understanding LangTest and testing out different test cases, and contributing to writing out deliverables and progress summary.
- **Focus Going Forward:** Trialing the dataset Ana found and maybe narrowing out with the team what we want to do going forward.
- **New Learning:** Improved my knowledge on datasets and ways to detect biases in LLMs such as BBQ, Amazon, LegalBench, NER. So many of these things I did not know before as well as playing with data using LangTest and HuggingFace which are both new interfaces I encountered from this project.
- **Area for Improvement:** Develop a better understanding of data evaluation and machine learning techniques.

Kennedy Martin:

- **Contributions:** Assisted with some administrative tasks (creating Github repository, setting up meeting invites & reminders). Researched bias detection and mitigation. Learned/practiced Hugging Face and Langtest. Used Google colab to manipulate the BBQ dataset. Also utilized Google colab to study sentiment analysis in regards to social/economic standing.

- **Focus Going Forward:** Researching how to create robust prompts for NER. Create a prompt for a NER.
- **New Learning:** HuggingFace, LangTest, improved working with JSON files, improved working with Github. Improved ability to read and understand research papers, improved communication skills in regards to asking the right questions.
- **Area for Improvement:** Improve my understanding of performance metrics, need to review past notes. Review Scikit learn and their metrics.

###