Capstone Project Progress Report: Synthetic Students

Matías Hoyl

December 13, 2024

1 Introduction

This progress report provides an update on the development of my capstone project, "Synthetic Students: Using Item Response Theory to Guide LLM-Based Answer Prediction." It outlines the progress made since the initial proposal, addresses the changes in direction, details the work completed during the Fall quarter, and presents a revised timeline for the Winter and Spring quarters.

1.1 Project Evolution

As a brief overview of my project and to give necessary context, it's important to mention that my original capstone proposal aimed to analyze a large-scale dataset of online tutoring sessions from the "Class Dojo Tutor" platform. However, due to logistical challenges in accessing and utilizing the Class Dojo dataset and given that I had access to a different dataset, I pivoted to a new direction that I believe to be more feasible and potentially more interesting.

1.2 Project Overview

My current project focuses on creating "synthetic students" using a combination of Item Response Theory (IRT) and Large Language Models (LLMs). The goal is to develop AI systems that can simulate how real students would respond to test questions, potentially aiding in the initial calibration of assessment items. This project leverages a dataset from Zapien, an educational technology platform, containing over 280,000 student interactions on mathematics questions. The core research questions are:

1. Can LLMs effectively simulate student response patterns when given information about student abilities and question characteristics?

2. What contextual information (such as prior performance, topic mastery, or demographic data) most effectively guides LLMs to generate realistic student responses?

This report details the progress made on these fronts, including data cleaning and preprocessing, exploratory data analysis, initial experimentation with LLMs, and the development of a platform to facilitate further experimentation.

2 Progress

2.1 Fall Quarter Objectives and Accomplishments

The primary objectives for the Fall quarter were centered around finding a viable research direction, preparing the dataset, conducting preliminary experiments, and outlining the project's structure. Here's a breakdown of the progress:

1. **Refining the Research Idea:** The initial phase involved extensive brainstorming and consultation with faculty, peers, and industry professionals. This process was crucial after pivoting from the original Class Dojo proposal. These discussions guided me toward leveraging the Zapien dataset, which contains rich information on student-question interactions. My conversations with professor Ben Domingue and Phd students Carrie Townley-Flores and Anya Ma were particularly helpful in refining my research idea.

2. **Data Preparation and Exploration:** A significant portion of the Fall quarter was dedicated to cleaning, preprocessing, and exploring the Zapien dataset. This involved removing irrelevant data, handling missing values, and creating new variables to enhance the dataset's utility for analysis. Exploratory data analysis was conducted to understand the distribution of key variables and identify initial patterns. This step was more time-consuming than initially anticipated but was necessary to ensure data quality. Also, having the opportunity to work with real-world data has been a valuable learning experience, and it has provided me with insights into the challenges of working with educational data.

3. **Literature Review and Conceptualization:** In parallel with data preparation, I conducted a literature review focusing on IRT, LLMs in education, assessment methods, prompting techniques, and the use of AI in educational settings. This research helped me identify a unique angle for my project, focusing on the intersection of IRT and LLMs to simulate student responses. I found that while LLMs have been used extensively in generating teacher-centric content and simulating learning scenarios, their application in directly simulating student responses remains relatively underdeveloped.

4. **Initial Experimentation and Platform Development:** I developed a platform to streamline interactions with LLM APIs (OpenAI, Anthropic, Google) and store experimental results systematically. This platform was instrumental in conducting preliminary experiments where LLMs were used to simulate student responses under different conditions. These experiments provided initial insights into the feasibility of the project and the effectiveness of different prompting strategies. Initially, I began working with Langchain and Langsmith for experimentation and evaluation of LLM workflows, but I quickly reached the limits of the free version. To avoid payment, I

built my own version of "Langsmith". This was time-consuming, but it was worth it to be able to conduct and review experiments more quickly and easily in the future.

5. **Report and Presentation Preparation:** The final stage of the Fall quarter involved compiling the findings and insights into a draft report and preparing a presentation to share the progress with the teaching team and peers. This process helped synthesize the work done and gather valuable feedback for the next steps. The feedback I received was very helpful and has given me a clearer direction for the Winter quarter.

2.2 Obstacles and Learnings

Several obstacles were encountered during the Fall quarter, each providing valuable learning opportunities:

1. **Data Cleaning and Preprocessing Challenges:** The Zapien dataset required extensive cleaning and preprocessing. Many variables were not immediately useful, and new features had to be engineered to support the analysis. This process was more time-consuming than initially estimated, delaying the start of experimentation. Learning: This experience underscored the importance of thorough data preparation in research projects and provided hands-on experience in data wrangling.

2. **Platform Development Time:** Building a custom platform for LLM experimentation took longer than expected. While tools like Langchain and Langsmith were initially helpful, their limitations necessitated the development of a bespoke solution. Learning: This challenge highlighted the trade-offs between using existing tools and developing custom solutions. The custom platform, while time-consuming to build, will enable more efficient experimentation moving forward. This will help me to iterate faster and test more hypotheses in the Winter quarter.

3. **Prompt Engineering Complexity:** Developing effective prompts for the LLMs to accurately simulate student behavior proved challenging. Translating IRT parameters into a language understandable by LLMs and ensuring they appropriately considered student context required multiple iterations. Learning: This highlighted the intricacies of prompt engineering and the need for a deep understanding of both the subject matter and the capabilities of LLMs. I learned that prompt engineering is an iterative process and that it requires a lot of experimentation to find the right prompt.

4. **Concerns about Result Robustness:** While initial experiments showed promising results, there's a concern that these findings might not hold under more rigorous testing or with larger sample sizes. Contingency: The priority for the Winter quarter is to conduct more extensive experiments, varying models and prompts to assess the robustness of the initial findings. If the effects are not sustained, alternative research angles within the synthetic student framework will be explored. This may include focusing on different aspects of student behavior or refining the IRT parameters used.

2.3 Project Timeline: Winter and Spring Quarters

Date	Task
January 15	Refine experimental design and conduct robustness tests with different LLM models and prompts.
January 22	Enhance data visualization techniques to better analyze synthetic student performance.
February 1	Develop additional evaluation metrics to measure alignment between LLM-generated and real student responses.
February 9	Submit second draft of the full report. Include refined methods and initial results.
February 16	Receive and incorporate peer feedback on the second draft.
February 22	Begin development of a prototype platform for synthetic student-based IRT parameter estimation.
March 1	Complete code submission, including platform updates and experiment workflows.
March 8	Conduct code review and address identified issues.
March 15	Finalize testing of synthetic students for unseen question sets and analyze IRT parameter correlation.
March 22	Submit final progress report and second draft to faculty advisor.
April 1	Design interactive visualizations for synthetic student performance insights.
April 10	Begin testing prototype platform with mock user data to simulate educator interaction.
April 25	Submit draft poster for peer and faculty review.
May 2	Incorporate written feedback on the draft poster.
May 9	Submit third draft of the full report to faculty advisor.
May 15	Develop user guide and documentation for synthetic student platform.
May 23	Submit final poster and final paper for grading.
Ongoing	Attend seminar sessions and participate in continuous project feedback and iteration.

3 Conclusion

Overall, I believe the project is on track to meet its objectives, albeit with a revised timeline and scope. The work completed in the Fall quarter has laid a solid foundation for the project. The initial experiments with LLMs are promising, and the custom platform will facilitate more extensive testing in the Winter quarter.

The revised timeline presented above outlines the key milestones for the Winter and Spring quarters, including expanded experimentation, development of a platform for external use, and preparation of the final report and presentation.

Support from the teaching team during the Winter quarter, particularly in reviewing

experimental designs and providing feedback on the interpretation of results, would be greatly appreciated. Additionally, any suggestions for further refining the research questions or exploring alternative approaches would be welcome.