EDUC259E Capstone Progress Report (Winter Quarter)

Matías Hoyl

1 Introduction

My capstone project currently focuses on predicting Item Response Theory (IRT) difficulty parameters for new assessment questions *without* needing to pre-test them on students. The core idea is to use a combination of question features, including linguistic characteristics and pedagogical insights automatically extracted using Large Language Models (LLMs), to train a model that simulates student response patterns. From these simulated patterns, we can then estimate the IRT difficulty.

2 Recap of Fall Quarter

As outlined in my Fall progress report, the initial phase involved pivoting from the original Class Dojo idea to using the Zapien dataset. The Fall quarter focused on data cleaning, exploration, and initial experiments simulating student responses using LLMs guided by basic IRT information. The plan heading into Winter was to conduct more rigorous testing on these LLM simulation methods to check if the promising early results held up.

3 Winter Quarter

The Winter quarter involved significant exploration and refinement of the project's methodology.

3.1 Testing the Initial LLM Simulation Idea

I started Winter by trying to make the initial LLM simulation approach (using IRT scores to guide LLMs) more robust. While the idea was interesting and showed some initial promise, further experiments with more repetitions and stricter testing revealed that the effects were either inconsistent or not strong enough to be reliable for predicting difficulty accurately. This was an important learning: early promising results need thorough validation.

3.2 Exploring Deep Knowledge Tracing

Realizing the limitations of the first approach, I briefly explored using Deep Knowledge Tracing (DKT) combined with transformer models (like ModernBERT). The idea was to model individual student knowledge over time based on their answer history. This seemed like a powerful way to capture learning dynamics. I even started training models for individual students. While this method was technically interesting and the models started showing decent performance (F1 scores > 0.75), it felt like it might be overly complex for the primary goal of estimating question difficulty, and it wasn't clear how easily the results from individual student models could be combined to get reliable IRT parameters for questions.

3.3 Pivoting to the Current Approach

Based on the learnings from the previous attempts, I shifted to the current approach which feels more direct and has yielded much stronger results. Instead of simulating individual answers one by one or modeling deep student knowledge, the focus is now on predicting the probability of a correct answer for any given student-question pair using a neural network.

- **Feature Engineering Focus:** A major part of the work this quarter involved creating features to feed into this model, including standard linguistic features (word counts, character counts), semantic embeddings using ModernBERT, LLM-extracted pedagogical features from Gemini Flash (steps to solve, mathematical skills, cognitive level, potential misconceptions), and user embeddings to represent student ability without letting it dominate the model.
- **Model Development:** I designed and trained a neural network that takes these features (user embedding + question features) and predicts the probability of a correct response.
- **IRT Estimation from Predictions:** The crucial step is using the trained model to predict outcomes for *all* users on *unseen* questions. These predicted response patterns are then fed into a 2PL IRT model to estimate the difficulty parameter for those unseen questions.

This current approach has shown encouraging results on the holdout test set (questions the model never saw during training). The estimated IRT difficulty parameters correlate strongly with the actual parameters (Pearson correlation of **0.85**, Spearman correlation of **0.96**). This suggests the method is effectively capturing question difficulty.

3.4 Obstacles and Learnings

1. **Initial Ideas Not Robust:** The biggest hurdle was realizing that the first two approaches (direct LLM simulation and DKT) weren't yielding the robust or directly applicable results needed for IRT parameter estimation. *Learning:* It's crucial to rigorously test initial findings and be willing to pivot if an approach isn't working as needed. Sometimes simpler, more direct methods are better.
2. **Feature Engineering Complexity:** Extracting meaningful features, especially using LLMs, was time-consuming. Ensuring the stability of LLM outputs required multiple runs and careful prompting (e.g., using Chain-of-Thought). *Learning:* High-quality features are essential for model performance. LLMs can provide valuable pedagogical insights, but require careful handling and validation.
3. **Balancing User vs. Question Influence:** Early attempts with simpler models (like LightGBM) struggled because user features dominated, making it hard to estimate question difficulty. *Learning:* The neural network architecture with user embeddings was key to properly balancing student ability and question characteristics. Model architecture choices matter significantly.
4. **Time Spent on Pivots:** Exploring multiple approaches took time, potentially delaying progress on the final method. *Learning:* While exploration is part of research, it's important to recognize when to settle on a promising direction. The pivots, however, provided valuable insights that informed the final, successful approach.

3.5 Time Estimate and Resources

I estimate I've spent approximately 15-20 hours per week on the capstone project this quarter. This time includes conducting numerous small experiments which, while not all successful, provided valuable insights that helped shape the project's direction. The support from Professor Domingue, feedback from peers during seminar sessions, and discussions with PhD students have been very helpful in navigating the research process and refining the project direction. The iterative nature of these small experiments was particularly instrumental in identifying promising paths to follow.

4 Conclusion

4.1 Overall Self-Assessment

Despite the pivots in methodology, I believe the project is now on track. The current approach of using a neural network with rich features to predict student responses and subsequently estimate IRT difficulty is yielding interesting results. The research journey, while not linear, was a necessary part of finding a method that truly works for the core research goal. I am confident that this direction will lead to a successful capstone project completion.

My main remaining concern is ensuring the generalizability of the findings, which I plan to address by potentially testing on another dataset (like ASSISTments) if time permits.

4.2 Revised Timeline for Spring Quarter

The focus for Spring is on finalizing the analysis, writing the paper, creating the poster, and preparing for the final presentation.

- **April 1:** Conduct ablation studies (testing feature importance). Refine analysis and visualizations.
- **April 10:** Refine the final paper draft (Introduction, Related Work, Methods).
- **April 17:** Polish remaining sections (Data, Results) and ensure coherence across paper.
- **April 25:** Submit draft poster for peer and faculty review.
- **May 2:** Incorporate written feedback on the draft poster. Continue writing paper (Discussion, Conclusion).
- **May 9:** Submit draft of the full report/paper to faculty advisor.
- **May 15:** Refine paper based on feedback. Prepare final presentation slides.
- **May 23:** Submit final poster and final paper for grading.
- **Ongoing:** Attend seminar sessions, incorporate feedback, practice presentation.

4.3 Support Requested

During the Spring quarter, I would appreciate feedback on:

- The interpretation and presentation of the results in the final paper.
- The structure and clarity of the final paper draft.
- Opportunities to practice the final presentation during seminar time.