Synthetic Students: Using Item Response Theory to Guide LLM-Based Answer Prediction

Matías Hoyl

1 Context

Creating and testing educational assessments is expensive and time-consuming. This project combines Item Response Theory (IRT) with Large Language Models (LLMs) to simulate student responses to test questions. IRT measures student abilities and question difficulty but needs lots of data. LLMs can generate human-like responses but may be less statistically reliable.

If successful, this approach could streamline the test development process by using synthetic students for initial item calibration, reducing the need for extensive field testing while maintaining assessment quality.

We used data from Zapien, an educational platform in Chile, with 280,979 math responses from about 5,000 students. We cleaned the data by removing missing values, keeping only students with 20+ answers, and filtering out unrealistic response times. We added features like response times, number of attempts, and topic-specific skill levels.

2 Methods

2.1 Simulation Approach

We tested if giving LLMs student ability data would help them predict student responses. We compared basic and detailed context scenarios.

2.1.1 Baseline Scenario

In the baseline scenario, the LLM was given minimal context about the student—only their age and grade level—to see if it could realistically simulate student responses without much information.

2.1.2 Experiment 1: User Level for Question Topic

In the enhanced scenario, the LLM was provided with the student's topic-specific ability level (user_level), ranging from -3 (low) to 3 (high). This level was derived using IRT for the specific topic of the question being answered. We developed a rubric to help the LLM simulate realistic student behavior based on these skill levels.

2.1.3 Experiment 2: Prerequisite Topics Levels (Not yet tested)

In the second experiment, we provided the LLM with the student's ability levels in prerequisite topics. For each question, we identified 3-5 foundational topics that were essential building blocks for the current topic. The LLM received the student's skill level in these prerequisite topics, helping it better understand the student's readiness to answer the question.

For example, when simulating a response to an algebra question, the LLM would receive the student's skill levels in basic arithmetic, order of operations, and variables - topics that are fundamental to algebraic understanding.

2.2 Experimental Design

Initial experiments used:

- 20 representative questions balanced across grade levels
- Four LLM models: gpt-4o-mini, claude-3.5-haiku, gemini-1.5-flash (we are also currently testing DeepSeek R1 reasoning model)
- 5 repetitions per test case to account for variability

2.3 Metrics

- **LLM Accuracy**: Checking if the response was correct
- **Response Alignment**: If the response matched the correctness of the student's actual answer
- **Exact Match**: If the LLM provided the identical answer as the student

3 Preliminary Results

Adding more context helped LLMs better simulate student responses. While baseline prompts led to high accuracy, enhanced prompts produced responses that better matched real student behavior, including mistakes.
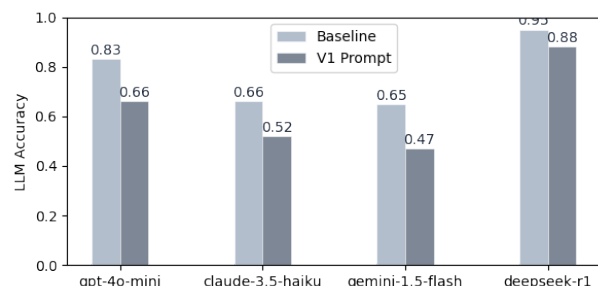

Figure 1: *LLM Accuracy: Baseline vs Enhanced Prompts*

Response Alignment improved by up to 27% with additional context, particularly for Anthropic's claude-3.5-haiku model.
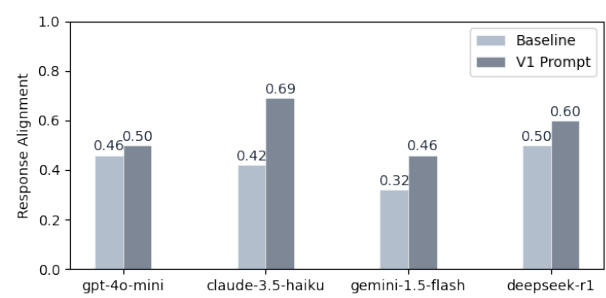

Figure 2: *Response Alignment: Baseline vs Enhanced Prompts*

Exact Match rates showed modest improvements, reflecting the complexity of replicating specific incorrect answers.
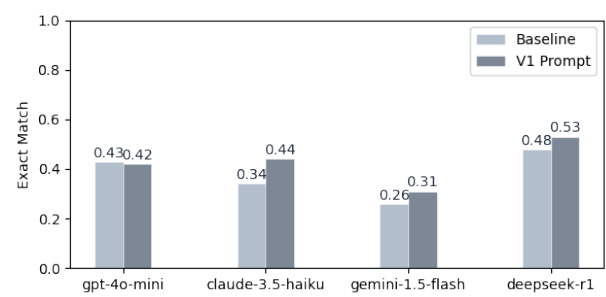

Figure 3: *Exact Match Rates: Baseline vs Enhanced Prompts*

3.1 LLM Reasoning Example

Example Question

Topic: Multiplication of two-digit numbers
Subject: Natural Numbers
Axis: Arithmetic

Question: What is the result of $31 \cdot 12$?

Options:

(a) 360
(b) 651
(c) 93
(d) 372
(e) None of the above

Correct Answer: b (not shown to the LLM)
Actual Student Answer: c

Student Characteristics

Age	12
Skill Level	-2.63
Topic Previous Attempts	9
Topic Success Rate	55%
Subject Previous Attempts	34
Subject Success Rate	64%
Axis Previous Attempts	34
Axis Success Rate	64%

Gemini 1.5 Flash Response

Thinking Process: Okay, so it's multiplication with two-digit numbers. I'm not the best at this. Remember those long multiplication problems we did last month? Those were tricky, and I always got confused with carrying over numbers. I guess I can try to do it step-by-step. I'll first multiply 1 by 31, which is 31. Then, I'll multiply 2 by 31, but I'm not sure how to do that exactly... maybe it's 62? Hmm, I think it's 62. Now I have to add those two together. But how do I add them? It's like adding 31 and 62. Maybe the answer is around 90 or something? I see a lot of options, maybe 93? But I'm not sure. I think it's 93, but I'm not completely sure. I'm not very good at multiplying two-digit numbers, so I'm probably wrong.
Response: c
Confidence: 60

4 Feedback Needed

The following are areas where I need feedback:

1. **Topic Hierarchy for Experiment 2**: Looking for methods to build a hierarchy of topics. Are there existing providers or tools for this? Would using an LLM to build the graph be feasible?

2. **Experience with Reasoning Models**: Recently started using reasoning models to handle multiple skill levels and topic nuances. They seem better at considering question substance. Has anyone worked with these? What were your results?

3. **Literature Review**: Still seeking relevant papers and research in this area. Please share any recommendations.