

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

Matías Hoyl

1 Overview

Educational assessment faces a significant challenge: developing and calibrating test items is expensive and time-consuming, requiring extensive student data. This project explores a novel solution by combining Item Response Theory (IRT) with Large Language Models (LLMs) to create "synthetic students" - AI systems that can simulate how real students would respond to test questions.

IRT provides robust, objective measurements of student abilities and question difficulty but requires substantial data for accurate calibration. LLMs, conversely, offer flexibility in understanding and generating human-like responses but may lack statistical rigor. By combining these approaches, we aim to leverage both IRT's psychometric precision and LLMs' adaptive capabilities to create more accurate student simulations.

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.

2 Data

The data for this study is derived from Zapien, an adaptive educational platform in Chile, which includes 280,979 responses from around 5,000 students on mathematics questions.

Key data preprocessing steps included the removal of missing values, filtering out students with fewer than 20 answers, and excluding unrealistic response times.

Features engineered for richer context included average response timestamps, cumulative attempts per student, and skill levels for specific topics.

3 Methods

3.1 Simulation Approach

The simulation approach aimed to provide the LLM with enough context, including IRT ability, so it could find sufficient signal in that context and replicate how a student with those characteristics would respond to questions. We developed baseline and enhanced context scenarios to test this.

3.1.1 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.

3.1.2 Enhanced Context Scenario

In the enhanced scenario, the LLM was provided with a richer context, including historical performance metrics and a psychometric measure of student ability (`user_level`) derived using IRT, ranging from -3 (low ability) to 3 (high ability). We developed an interpretative rubric to guide the LLM in simulating realistic student behavior based on these skill levels. The enriched prompt included prior successes, question difficulty, and other relevant academic details, helping the LLM generate more contextually accurate responses.

3.2 Experimental Design

Initial experiments used:

- 20 representative questions balanced across grade levels
- Three LLM models: gpt-4o-mini, claude-3.5-haiku, gemini-1.5-flash
- 5 repetitions per test case to account for variability

3.3 Metrics

We used three metrics to evaluate the performance of the LLM:

- **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

4 Preliminary Results

The preliminary results showed that providing LLMs with enhanced contextual information improved their ability to simulate student responses more realistically. Accuracy rates were high in baseline scenarios, indicating a tendency towards correct answers. However, when richer context was provided, LLMs generated responses that were more aligned with the correct/incorrect nature of actual student answers, indicating improved behavioral simulation.

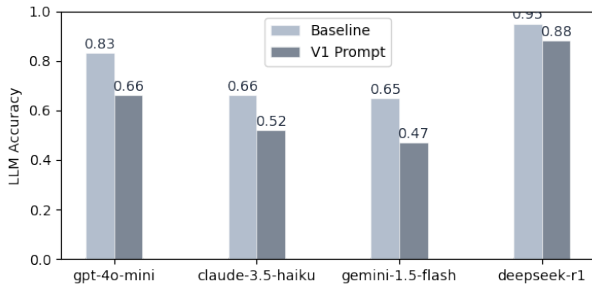


Figure 1: LLM Accuracy: Baseline vs Enhanced Prompts

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

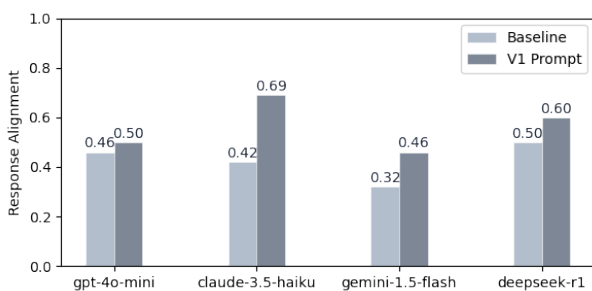


Figure 2: Response Alignment: Baseline vs Enhanced Prompts

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

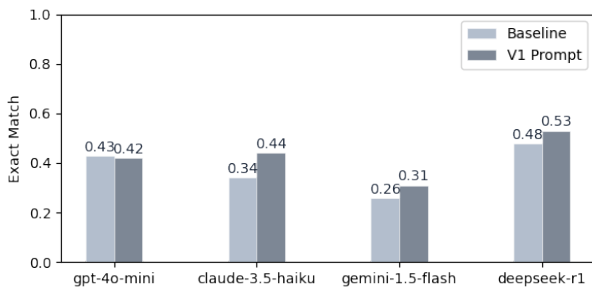


Figure 3: Exact Match Rates: Baseline vs Enhanced Prompts

4.1 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

5 Challenges

The following are specific areas where I would appreciate feedback:

- Statistical Significance of LLM Tests:** How can I ensure statistical significance in LLM experiments? What tests or methods should I use to validate my findings?
- Monitoring Soft Metrics:** Besides accuracy and exact match, should I track metrics like "believability" of responses or reasoning consistency? How can I evaluate these qualitative aspects?
- Literature Gaps:** Are there any studies similar to this work on simulating student behavior with LLMs?