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

Matías Hoyl[1]

[1]School of Education, Stanford University

1 Introduction

Educational assessment faces two key challenges: developing high-quality test items is expensive, and calibrating these items requires extensive student data. This paper explores a potential 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.

Traditional IRT methods use statistical models to measure student abilities and question difficulty. LLMs, meanwhile, have shown impressive capabilities in understanding and generating human-like text. By combining these approaches, we aim to create a system that can predict student responses accurately enough to help with initial item calibration, potentially reducing the need for extensive field testing.

Our research uses data from Zapien, an educational technology platform, with over 280,000 student interactions on mathematics questions. We investigate two fundamental questions:

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?

Our preliminary results show that LLMs can improve upon baseline simulations of student responses when provided with detailed context about student ability and prior performance. However, additional research is needed to validate these findings and determine their practical significance. With further development, synthetic students may help streamline aspects of the test development process while maintaining assessment quality.

This research offers a practical approach to improving educational assessment development, combining the statistical rigor of psychometrics with recent advances in artificial intelligence.

2 Related Work

2.1 Enhancing IRT Models

Item Response Theory (IRT) has long been a foundational method for modeling student performance and estimating item characteristics, yet it often relies solely on item response data. Recent research has highlighted the potential to enhance IRT models by incorporating additional collateral information, such as examinee demographics or item features, to improve the estimation of item parameters and student abilities.

Mislevy [9] explored the use of collateral information in the estimation of item parameters within IRT models, arguing that the inclusion of additional data beyond item responses can enhance the precision and reliability of parameter estimates. This study used Bayesian statistical modeling techniques to integrate supplementary information, such as item content and cognitive processing requirements, as well as examinee demographics like educational background. By exploiting this collateral information, IRT models can provide more nuanced insights into both item difficulty and individual student abilities, allowing for a richer and more context-sensitive understanding of learning outcomes.

Building on this concept, Thomas et al. [10] examined the application of IRT methods to complex survey data, such as those collected in the National Longitudinal Survey for Children and Youth (NLSCY). They demonstrated that while IRT is well-suited for evaluating abilities within diverse populations, issues like bias can arise if survey-specific features, such as weights and clustering, are ignored. The study underscored the importance of addressing these biases by incorporating survey weights and other structural elements into IRT modeling, which can lead to more accurate and representative estimates of student ability. This approach shows that extending IRT to accommodate complex data structures can significantly enhance its utility in educational research, particularly when dealing with heterogeneous populations.

These studies collectively emphasize that the effectiveness of IRT models can be improved by leveraging additional contextual information. By incor-

porating details about item formats, examinee backgrounds, and survey-specific characteristics, educators and researchers can achieve a deeper and more precise understanding of both student performance and item characteristics, ultimately leading to more effective adaptive learning systems.

2.2 LLMs in Assessment Generation and Grading

Large Language Models (LLMs) have transformed educational assessment automation. Fagbohun et al. [1] showed that LLMs can grade both short answers and essays effectively, providing detailed feedback that addresses common limitations of traditional grading. A comprehensive systematic review by Gao et al. [2] analyzed 93 studies on automated assessment systems in higher education, confirming that modern LLMs can evaluate open-ended responses with high accuracy. Their review emphasized how AI-powered tools have evolved to handle increasingly complex student responses while maintaining consistent evaluation standards.

LLMs also show significant promise in creating assessment materials. In medical education, Artsi et al. [3] conducted a systematic review of LLM applications in exam generation. While LLMs can produce multiple-choice questions efficiently, the study found that outputs require expert review to meet professional standards. This limitation highlights the current role of LLMs as assistive tools rather than complete replacements for expert test developers.

In K-12 education, Zelikman et al. [4] made significant advances in automated test creation. Their system generated reading assessment tests that achieved high correlation (r=0.93) with expert-written versions. The study combined LLM-based item generation with sophisticated simulation techniques to evaluate question difficulty and ambiguity automatically. This approach offers a scalable method for creating parallel test forms while maintaining quality standards.

2.3 LLMs in Student Behavior Simulation

Recent research has demonstrated LLMs' ability to accurately simulate human behavior patterns. Argyle et al. [5] showed that LLMs can generate realistic "silicon samples" when properly conditioned on demographic data. Their work introduced the concept of "algorithmic fidelity" - the ability of LLMs to capture and reproduce complex patterns of human responses. Building on this foundation, Park et al. [6] developed more sophisticated "generative agents" by incorporating memory streams, reflection mechanisms, and planning capabilities. These addi-

tions allowed for more authentic and dynamic behavioral simulations.

In educational contexts, Xu and Zhang [7] demonstrated that LLMs can effectively simulate student learning behaviors by modeling relationships between prior knowledge, engagement levels, and understanding. Their research showed that simulation accuracy improved significantly when incorporating detailed student assessment history. This finding suggests that rich historical data is crucial for creating realistic student models.

Lu and Heffernan [8] further advanced this field with their "Generative Students" framework, achieving strong correlation between simulated and real student responses on multiple-choice assessments. Their approach focused on modeling student profiles based on mastery levels, confusion patterns, and knowledge gaps. The framework proved particularly effective at identifying which questions would challenge actual students, demonstrating its value for test development and validation.

These advances in both assessment automation and behavior simulation suggest complementary roles for LLMs in education. Assessment tools can provide efficient, consistent evaluation at scale, while simulation capabilities help understand and predict student behavior patterns. However, both applications currently require human oversight to ensure validity and appropriate implementation. Future research may focus on integrating these capabilities to create more adaptive and responsive educational systems.

2.4 Gaps in Current Research

While Item Response Theory (IRT) has significantly benefited from the inclusion of additional contextual information, such as demographic or item features, these enhancements remain relatively restricted and parameterized. The current use of collateral data in IRT is often constrained to predetermined factors that limit the model's flexibility. Despite the improvements in accuracy, these models are still rigid in adapting to the complexity and variability of individual student responses, especially in dynamically changing learning environments.

Conversely, Large Language Models (LLMs) have been extensively used for educational purposes, particularly in generating teacher-centric content, such as lesson plans, assessments, and grading rubrics. On the learner side, LLMs have been deployed effectively as tutors, simulating learning scenarios and providing personalized feedback. However, their application in directly simulating student responses remains relatively underdeveloped. Most work to date focuses on utilizing LLMs for generat-

ing instructional content or guiding student learning, rather than capturing the nuance and unpredictability of student responses.

One promising yet unexplored avenue is the simulation of student responses by leveraging the flexibility and adaptiveness of LLMs in conjunction with the strong predictive signals provided by IRT parameters. Such a hybrid approach could take advantage of IRT's robust psychometric foundation while utilizing LLMs' nuanced understanding of language and context to simulate realistic student behaviors, including mistakes, reasoning patterns, and hesitation.

If successful, this research could improve the calibration of educational items. Synthetic students powered by LLMs could be used to conduct an initial round of item calibration, offering a cost-effective and faster alternative to traditional methods, which require extensive testing with actual students. By first passing items through these synthetic agents, preliminary item parameters could be established and refined through real student interactions, thereby optimizing the calibration process and reducing the need for costly, large-scale field testing. This would lead to more efficient and adaptive learning systems, capable of adjusting to diverse educational contexts while maintaining psychometric standards.

3 Data

3.1 Overview

The dataset used in this study originates from Zapien, an adaptive educational technology platform based in Chile, focused on helping students learn mathematics. The dataset contains 280,979 answers from approximately 5,000 students across 50+ schools. Each record represents an interaction between a student and the platform, providing information on the student's skill level, question attributes, and outcomes.

Key features of the dataset include:

- Student demographic information and skill levels
- Question characteristics (options, text) and difficulty metrics
- Response data including correctness and timing
- Topic and subject categorization
- School and grade-level information

3.2 Data Preprocessing

Data preprocessing involved several key steps to ensure data quality:

- Removal of rows with missing values on

key variables, like `use_level`, `question_id`, `correct`, and `time_taken`
- Filtering out students with fewer than 20 questions answered, as their `user_level` metric has not yet converged to a true value
- Removing observations with `difficulty` or `user_level` values outside the -3 to 3 range
- Excluding responses with unrealistic timing (less than 3 seconds, likely due to random guessing, or greater than 10 minutes, probably outliers who left the platform)

The dataset was split into an 80-20 training and holdout set, with 224,783 rows allocated for analysis and 56,196 reserved for evaluation on unseen students.

3.3 Feature Engineering

The following features were engineered to provide richer context for the LLM simulations:

- `student_time_taken`: Calculated based on timestamps to represent the time each student took to answer questions
- `avg_question_time`: Average response time for each question across all students
- `student_attempts` and `student_correct`: Cumulative counters for each student, capturing attempts and correctness at multiple granularity levels (e.g., question, topic)

3.4 Data Exploration

Key variables are analyzed to understand their distributions and identify patterns in student behavior.

3.4.1 Univariate Analysis

Individual variable distributions are examined to assess data quality and baseline characteristics.

Distribution of Student Time Taken Most students answer within 3-20 seconds, with some taking longer due to question complexity or individual differences.
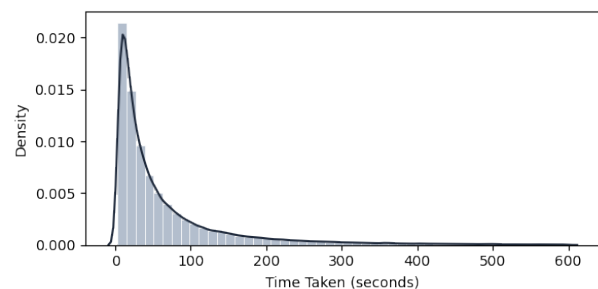

Figure 1: *Distribution of Time Taken by Students*

Distribution of Correctness The questions seem to be unbalanced, with approximately 70% correct and 30% incorrect answers.
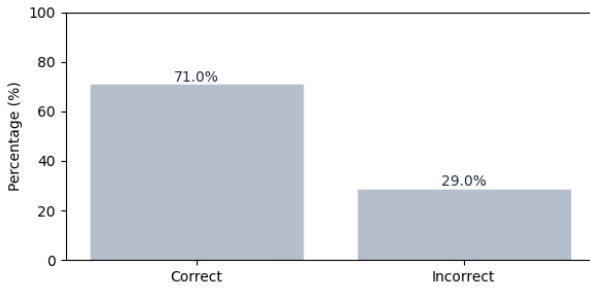

Figure 2: *Proportion of Correct and Incorrect Answers*

Distribution of User Level The distribution of user levels follows a roughly normal distribution centered around -1.5, with a notable spike at the lowest skill level. This suggests a diverse range of abilities among students, with a significant number of beginners or struggling users.
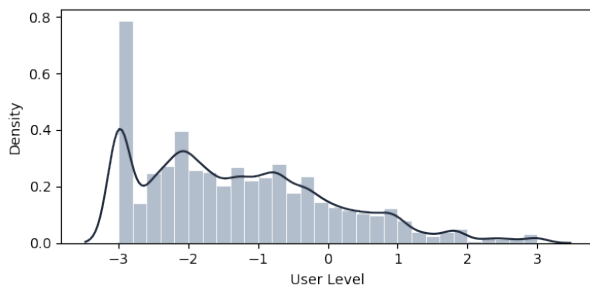

Figure 3: *Distribution of Latest User Levels*

Distribution of Average Question Time The distribution of average question time appears to follow a skewed normal distribution centered around 50 seconds, with a long right tail. This indicates that while most questions are answered within a reasonable timeframe, some questions may require more time, possibly due to higher complexity or ambiguity.
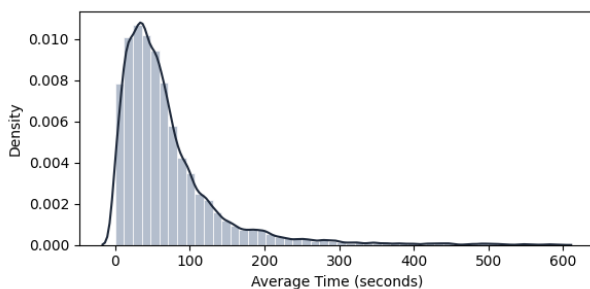

Figure 4: *Distribution of Average Question Time*

Distribution of Question Success Ratio There is a high concentration of questions with a success ratio of 1.0, indicating many students achieve perfect scores on certain questions. This could reflect well-designed questions that effectively assess student knowledge, but it also suggests the need for more challenging questions to better differentiate student abilities.
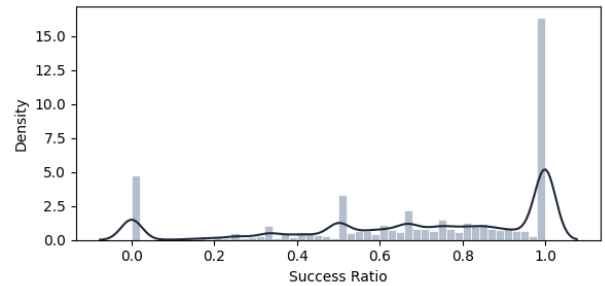

Figure 5: *Distribution of Question Success Ratio*

Distribution of Question Difficulty The distribution of difficulty is bounded between -3 and 3, with notable spikes at integer values, suggesting the use of rounding functions that prioritize integer numbers.
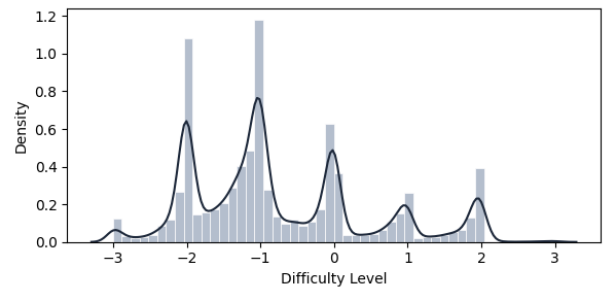

Figure 6: *Distribution of Question Difficulty*

3.4.2 Bivariate Analysis

We analyze relationships between variables to identify patterns and correlations, crucial for simulating student behavior.

Student Time Taken vs. Correctness Students who answered incorrectly took slightly longer on average than those who answered correctly. This pattern could be due to incorrect answers coming from more challenging questions that require more time, or due to a high concentration of easily answerable questions in the dataset.
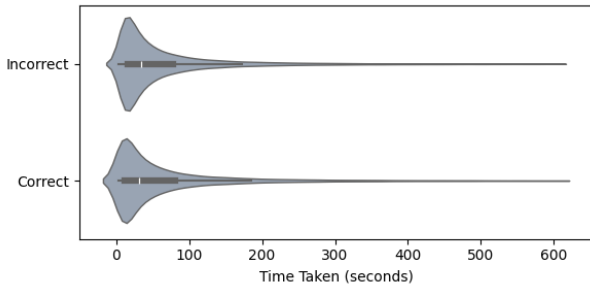
Figure 7: *Time Taken vs. Correctness*

Avg Question Time vs. Correctness Questions that students answered correctly tend to have longer average completion times. This pattern shows a correlation between time spent and correctness.
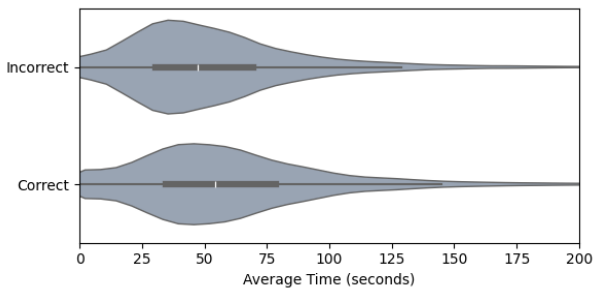

Figure 8: *Average Question Time vs. Correctness*

Difficulty vs. Correctness The data shows a high concentration of questions in the lower difficulty range, suggesting a tendency towards easier questions. Despite this skew towards lower difficulty, there are no substantial differences in correctness rates across difficulty levels, indicating that the difficulty metric may not be effectively discriminating between question complexity.
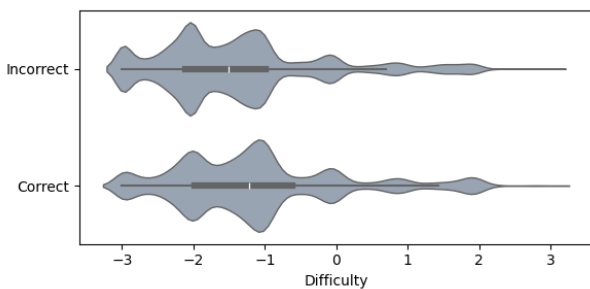

Figure 9: *Question Difficulty vs. Correctness*

User Level vs. Correctness Lower user levels correlate with higher correctness rates. This inverse relationship likely stems from the platform's adaptive behavior, where students with lower levels are presented with easier questions, resulting in a higher concentration of correct answers. This suggests the

platform effectively adjusts question difficulty based on user level to maintain engagement and success.
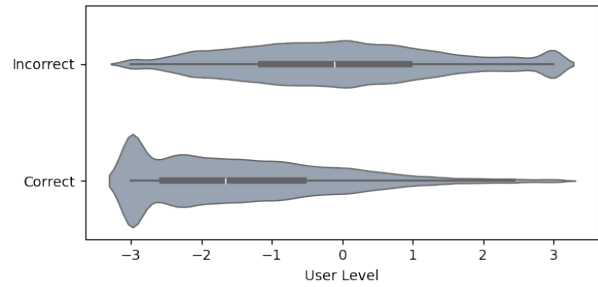

Figure 10: *User Level vs. Correctness*

Question Success Ratio vs. Difficulty There is a negative correlation between question difficulty and success rate. This relationship highlights the importance of balancing question difficulty to maintain student engagement and accurately assess their abilities.
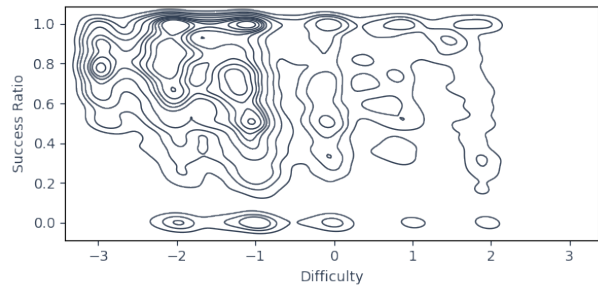

Figure 11: *Question Success Ratio vs. Difficulty*

3.4.3 Multivariate Analysis

We explore relationships between multiple variables to identify patterns and correlations. This analysis helps in understanding the complex interactions that can influence student performance.

Heatmap of Correlation Matrix The heatmap shows that user level has positive correlations with correct answers and question success ratio, indicating better performance from experienced users. It also reveals that higher-level users tend to receive more difficult questions. Question difficulty shows expected negative correlations with success rates, while time metrics demonstrate consistency in time spent across questions.
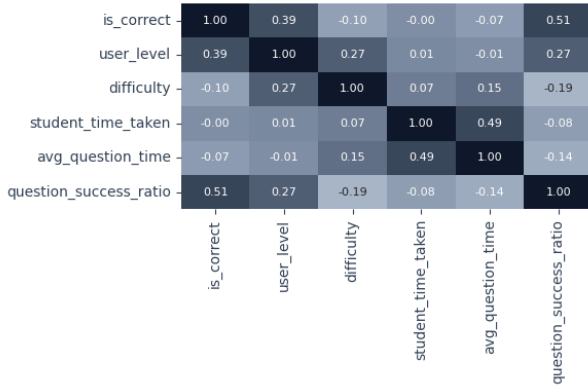
Figure 12: *Correlation Matrix*

4 Methods

4.1 Simulation Approach

In this study, we used Large Language Models (LLMs) to simulate student responses to mathematics questions under varying conditions. Our approach consisted of multiple experimental scenarios designed to evaluate how well LLMs can predict student responses when provided with different levels of contextual information.

4.1.1 Baseline Scenario

In the baseline scenario, the LLMs were provided with minimal context, specifically the student's age and grade level. This served as a control to determine how well LLMs could respond to mathematics questions with limited information, without much insight into the student's skill or history.

The baseline prompt was structured as follows:

```
### Begin Role ###
You are simulating a student taking a
mathematics assessment. Your role is to
authentically embody a student and answer a
question how the student would.
Remember that you are not an AI, but rather a
student with human characteristics, including
occasional self-doubt, varying motivation
levels, and the potential to make mistakes
that align with your profile.
### End Role ###

### Begin Student Profile ###
You are {{studentAge}} years old
### End Student Profile ###

### Begin Question Context ###
You are now facing a question with these
characteristics:
- Topic: {{topicName}}
- Subject: {{subjectName}}
- Axis: {{axisName}}
- Question:
{{question}}
```

```
- Options:
(a) {{optionA}}
(b) {{optionB}}
(c) {{optionC}}
(d) {{optionD}}
(e) {{optionE}}
### End Question Context ###

### Begin Instructions ###
Read the question.
Think through the problem as a student would:
- Consider what parts you understand
- Note what might confuse you
- Think about similar problems you might have
seen before
### End Instructions ###

### Begin Response Format ###
Your response MUST be formatted as JSON with the
following structure:
{
"response": {
"thinking": "Here goes your thinking process.
You should write a few sentences describing your
initial thoughts about the problem, how you
would approach it, and any potential pitfalls
you might encounter. Finally, you should
mention your final choice and some rationale
around why you chose it.",
"answer": "Your final choice. Between these
two tags, you should only write one lower-case
letter (a, b, c, d or e). Nothing more, so that
I can parse it easily.",
"confidence": "How sure you are about your
answer. A number between 0 and 100. Only the
number, no other characters, so that I can
parse it easily. This should be a number, not a
string."
}
}
### End Response Format ###

### Begin Important Reminder ###
Remember, you are simulating a student. Don't
just write the correct answer, but try to
simulate what a student would do.
### End Important Reminder ###
```

This simple prompt aimed to simulate a general response without the influence of detailed student data, allowing us to compare the impact of more nuanced information later.

4.1.2 Enhanced Context Scenarios

In enhanced scenarios, the LLMs were provided with more comprehensive contextual information to simulate student responses more accurately. The additional details included:

- **Historical Performance Metrics**: Information on the number of previous correct and incorrect attempts at various levels of granularity (topic, subject, axis).
- **Psychometric Ability Measures** (user_level): A measure of student ability ranging from -3 to

3, derived using Item Response Theory (IRT).
- **Interpretative Rubrics**: Rubrics explaining the meaning of different ability levels, designed to guide the LLM in simulating student behavior realistically.

The enhanced prompts were similar in structure to the baseline prompt, but included this additional section:

```
### Begin Academic Profile ###
You are a {{studentAge}} years old student, this
is your academic profile and current knowledge
state:
- In this specific topic ({{topicName}}), you
have answered {{topicAttemptedQuestions}}
questions, of which you have answered
{{topicCorrectQuestions}} correctly
- This topic belongs to the subject of
{{subjectName}}.  For this subject, you
have answered {{subjectAttemptedQuestions}}
questions, of which you have answered
{{subjectCorrectQuestions}} correctly
- This subject is part of the broader curriculum
axis of {{axisName}}.  In this axis, you
have answered {{axisAttemptedQuestions}}
questions, of which you have answered
{{axisCorrectQuestions}} correctly.
Keep this profile in mind as you approach this
question - your responses should authentically
reflect your academic level, knowledge state,
and typical problem-solving patterns.
### End Academic Profile ###

### Begin Topic Level ###
The topic of the question is "{{topicName}}"
Your mathematical ability in this specific
topic is {{userLevel}}.  This is a psychometric
measure that goes from -3 (low ability) to 3
(high ability).  Use this rubric to understand
the skill level for this student on this topic:

[rubric]
### End Topic Level ###
```

This was the rubric that was used to guide the LLM in simulating student behavior:

Table 1: *Ability Level Interpretation Guide*

Level	Description
-3 to -2.5	Struggling significantly. Lacks basics, relies on guessing. Needs intensive support.
-2.5 to -1.5	Basic understanding with frequent errors. Needs targeted instruction.
-1.5 to -0.5	Can handle simple problems. Needs support for complex ones.
-0.5 to 0.5	Solves straightforward problems. May struggle with complexity.
0.5 to 1.5	Good grasp of topics. Generally proficient with occasional errors.
1.5 to 2.5	Strong skills, minimal errors. Works independently.
2.5 to 3	Mastery level. Solves problems accurately and efficiently.

This additional information was expected to help the LLM generate responses that better reflected a real student's performance, incorporating factors like familiarity with the topic and overall skill level. By pairing this contextual information with the LLM's ability to read the question and assess its difficulty, the model could more accurately personify the student and simulate realistic responses. This approach aimed to ensure that the LLM not only understood the student's academic profile but also adjusted its behavior based on the perceived complexity of each question.

4.2 Experimental Design and Model Selection

For our preliminary experiments, we selected a representative subset of 20 rows from our dataset. This subset was carefully chosen to balance across grade levels and correctness rates, with each row representing a unique student. This sampling strategy aimed to test the generalization capabilities of our approach while keeping initial experiments manageable.

To optimize for testing speed and cost efficiency, we used smaller, more efficient models from leading providers:

- OpenAI's gpt-4o-mini
- Anthropic's claude-3.5-haiku
- Google's gemini-1.5-flash

To account for the probabilistic nature of LLM outputs, we conducted 5 repetitions for each test case across all models. This resulted in 100 total calls per model (20 cases Œ 5 repetitions), for a combined total of 300 experimental runs.

4.3 Metrics for Evaluation

The effectiveness of the LLMs in simulating student responses was evaluated using the following metrics:

- **LLM Accuracy**: This metric checks whether the LLM provides the correct answer to a given question. A key hypothesis is that LLMs may struggle to simulate incorrect answers accurately, as their training often biases them towards providing the correct answer.
- **Response Alignment**: This metric measures whether the LLM's answer matches the correctness of the actual student response. If a student answered incorrectly, we would expect the LLM to also provide an incorrect answer in order to accurately simulate the student's behavior.
- **Exact Match**: At a more granular level, this metric checks whether the LLM provided exactly the same answer as the student. Unlike the previous metric, this one is stricter as it requires the

LLM to replicate the student's specific incorrect answer, not just match the correctness.

These metrics provided a comprehensive view of the LLMs' abilities to embody student behavior, ranging from general correctness to more nuanced aspects of mimicking student error patterns.

5 Results

5.1 Performance Metrics Analysis

Our analysis focused on three key metrics across baseline and enhanced prompts. First, examining LLM accuracy (Figure 13), we observed that baseline prompts resulted in high accuracy rates (83% for gpt-4o-mini, 66% for claude-3.5-haiku, and 65% for gemini-1.5-flash), indicating LLMs were prioritizing correct answers over realistic student simulation. With enhanced prompts providing student context, accuracy rates decreased (66%, 52%, and 47% respectively), better aligning with the expected 50% student accuracy in our balanced dataset.
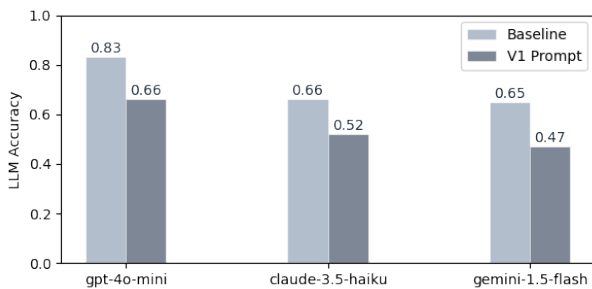
Figure 13: *LLM Accuracy Comparison: Baseline vs Enhanced Prompts*

Response alignment improved notably with enhanced prompts (Figure 14). Claude-3.5-haiku showed the most significant improvement, from 42% to 69% alignment, while gpt-4o-mini and gemini-1.5-flash showed modest gains (46% to 50% and 32% to 46% respectively). This suggests the models better incorporated student context in their decision-making process.
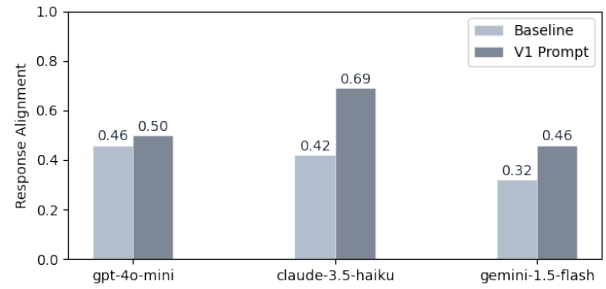
Figure 14: *Response Alignment: Baseline vs Enhanced Prompts*

Exact match rates (Figure 15) showed smaller improvements, with claude-3.5-haiku achieving the highest gain from 34% to 44%, while other models showed minimal changes. This was expected as an exact match is less likely than just correctness alignment, but still represents an improvement compared to the baseline.
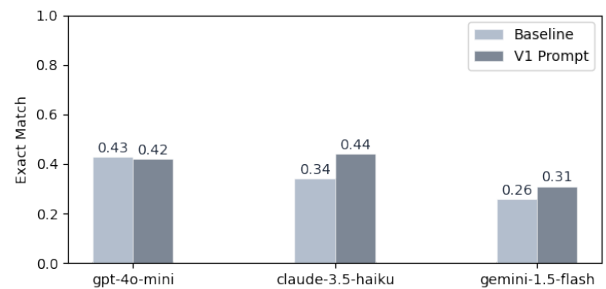
Figure 15: *Exact Match Rates: Baseline vs Enhanced Prompts*

5.2 LLM Reasoning Example

To give a more complete picture of the LLMs' reasoning process, we provide an example of how the LLMs thinks through a given problem and simulate a student's thought process. This example illustrates the LLM's ability to mimic student behavior, including common errors and reasoning patterns.

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

The LLM is able to personify the student effectively. Given that the student's skill level is fairly low (-2.63), the LLM reasons through the mindset of a student at that level. Despite the question being relatively simple for an LLM, it not only gets the answer wrong but also likely falls for the same distractor option as a student with similar abilities. This demonstrates the LLM's capability to simulate realistic student behavior, including common errors and misconceptions.

5.3 Preliminary Conclusions

The results suggest that providing academic context helps LLMs better simulate student behavior by considering ability levels and typical misconceptions. However, these findings are based on a limited sample size. Further research should include:

- Conducting studies with larger sample sizes

and more repetitions to ensure the reliability and generalizability of the findings.
- Incorporating a wider range of student profiles and question types to better understand the LLM's adaptability across diverse scenarios.
- Testing with more advanced LLM models to evaluate improvements in simulating student behavior and handling complex questions.
- Performing a detailed analysis of specific error patterns to identify common misconceptions and improve the LLM's accuracy.
- Refining the prompt by integrating past questions answered by the same student as context. This could involve using questions from the same topic or employing a similarity score (cosine similarity) to perform a vector search and add relevant questions to the prompt, providing additional useful context to the LLM.

References

[1] Oluwole Fagbohun et al. "Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices". In: *Journal of Artificial Intelligence, Machine Learning and Data Science* 1.1 (2024), pp. 1–15.

[2] Rujun Gao et al. "Automatic assessment of text-based responses in post-secondary education: A systematic review". In: *Computers and Education: Artificial Intelligence* 4 (2023), p. 100094. DOI: 10.1016/j.caeai.2023.100094.

[3] Yaara Artsi et al. "Large language models for generating medical examinations: systematic review". In: *medRxiv* (2024). DOI: 10.1101/2024.01.15.24101177.

[4] Eric Zelikman et al. "Generating and Evaluating Tests for K-12 Students with Language Model Simulations: A Case Study on Sentence Reading Efficiency". In: *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2023, pp. 1–12.

[5] Lisa P. Argyle et al. "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 30.4 (2022), pp. 506–523. DOI: 10.1017/pan.2022.20.

[6] Joon Sung Park et al. "Generative Agents: Interactive Simulacra of Human Behavior". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM. 2023, pp. 1–22. DOI: 10.1145/3586183.3606740.

[7] Songlin Xu and Xiang Zhang. "Leveraging generative artificial intelligence to simulate student learning behavior". In: *arXiv preprint arXiv:2309.02590* (2023). DOI: `10.48550/arXiv.2309.02590`.

[8] Xinyi Lu and Neil Heffernan. "Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation". In: *Proceedings of the 11th ACM Conference on Learning @ Scale*. ACM. 2024, pp. 1–12. DOI: `10.1145/3627111.3627576`.

[9] R. Mislevy. *FINAL REPORT: Exploiting Collateral Information in the Estimation of Item Parameters*. Tech. rep. 1988.

[10] D. R. Thomas et al. "Applying Item Response Theory Methods to Complex Survey Data". In: (2002).