Take Out Your Calculators: Estimating the Real Difficulty of Math Word Problems with LLM Student Simulations

Anonymous ACL submission

Abstract

Math word problems used in testing are usually piloted with human subjects to establish the item difficulty and detect differential item function. However these pilots are costly, thus for a need for a less costly approach that evaluates these questions. We show that large-language models to an extent can serve as a valuable first check, to help test developers effectively measure students' skills on a given subject matter. We do this by prompting Large Language Models(LLMs) to role-play Below Basic, Basic, Proficient, and Advanced 4th, 8th and 12th-grade students. We also add first names to simulate a more realistic classroom whose aggregate correct/wrong rate serves as a proxy for estimating question difficulty. We observe correlations between 0.61-0.76 with real student performance, where larger simulated classrooms (300 vs. 50 students) and the use of diverse student names resulted in better correlations than simulating only skill level with no names or using a single name with specified skill level for a classroom.

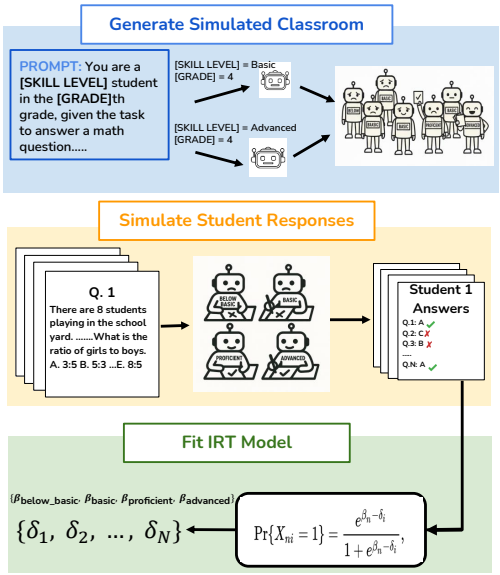

Figure 1: Top: We simulate our classrooms with LLMS roleplaying students at different skill levels based on specified skill distributions. Middle: Simulated students with varying abilities (represented by different skill levels) respond to question items. Bottom: We fit an IRT model to the simulated response data to estimate student ability parameters, β and item difficulty, δ.

1 Introduction

Multiple Choice Maths Problems are fundamental instruments of student evaluation and instruction. Because word problems test a student's ability to connect mathematical concepts to real-world scenarios, these items can interact in non-trivial ways with a student's knowledge and understanding of real-world concepts, independent of mathematical facility (Chipman et al., 1991).

Thus, there is a need for rigorous evaluation of test items which includes the careful, subjective cognitive analysis or modeling of question items by experts (Lei, 2007; Wu et al., 2025). Other evaluation methods involve the test taker in generating these question items (Singh et al., 2021) or rely on their retrospective performance data after psychometric-based analysis(Harris, 1989; Bond and Fox, 2013). A single pilot study can cost tens of thousands of dollars and take weeks to complete. Question item difficulty estimation provides educators and researchers, a way to validate that new curated question items are able to discriminate between students with varying ability levels, in order to reliably evaluate a students expertise on a given subject. Estimation frameworks like Item Response theory (IRT) offers powerful psychometric tools for objective difficulty estimation and extends beyond simple question text validity to enable the explicit identification of both difficulty and discriminability parameters of question items, however, it requires individual student response data which obtaining is challenging and time-consuming (Zelikman et al., 2023). Even when student response datasets exist,

access remains a major issue. These datasets are typically proprietary, and publicly available alternatives usually provide only aggregate student performance metrics rather than the individual response patterns that IRT requires for accurate parameter estimation, thus a need to simulate high-quality, individual student responses that are consistent with real-world data(Srivastava and Goodman, 2021; Scarlatos et al., 2025).

Recent works show LLMs can act as reliable 'silicon' subjects, reproducing human heuristics and behavioral patterns across trust tasks and other domains. (Xie et al., 2024; Argyle et al., 2023; Dillion et al., 2023; Manning et al., 2024; Yang et al., 2024). Prior studies have already sketched this direction in item difficulty estimation with simulated students: Generative-Student profiles built from knowledge components detect hard items without real data (Lu and Wang, 2024), the Classroom Simulacra framework which models class dynamics (Xu et al., 2025b), and GPT-based open-ended knowledge tracing produces realistic student answers that reveal mastery gaps (Liu et al., 2023). We present different prompt styles for simulating diverse student profiles: with varying skill levels of (Below Basic, Basic, Proficient, Advanced) and demographic name attributes and ensure the quality of the simulated responses by comparing against real world student performance. Our approach grounds these simulations against real student performance data using standardized psychometric techniques, validating their predictive power in estimating item difficulty accurately as shown in Figure 1. This provides test developers with insights for improving assessments proactively. Specifically, we address the following research questions:

1. Can open source LLMS reproduce real-world student performance and associated difficulty across varying skill levels?

2. How do different prompting strategies affect the alignment between simulated and actual student performance?

3. Can item difficulty estimates obtained from Rasch modeling of LLM-simulated answers reliably mirror those provided by test developers?

Difficulty	Total #	Grade 4	Grade 8	Grade 12
Easy	254	102	108	44
Medium	217	76	106	35
Hard	160	50	68	42
Total	**631**	**228**	**282**	**121**

Table 1: Breakdown of question difficulty by grade.

2 Preliminaries

2.1 Data

We collect 631 Multiple Choice Math Questions from the National Report Card website (National Center for Education Statistics (NCES)) for grades 4, 8 and 12. NAEP is a congress authorized project of the National Center for Education Statistics (NCES) with the Institute of Education Sciences of the U.S. Department of Education. These question items provide actual student performance statistics across the United States, providing a benchmark against which our LLM simulations can be evaluated. This data also serves as the only nationally representative and continuing assessment of student achievement in the United States. Questions from the NAEP have undergone rigorous development and validation processes. Each problem includes a multiple choice question, several answer choices, a correct answer, and metadata including difficulty, content area, grade level, and student performance statistics across demographic groups, including gender. Table 1 summarizes the distribution of question items by grade level and difficulty.

Despite the size of the data, the selected question items cover a diverse range of mathematical concepts. Figure 4 in the appendix presents the distribution of question items across content areas. Given that we are not conducting combinatorial experiments across multiple variables, but rather evaluating LLM capabilities against human benchmarks, this sample size provides sufficient coverage of key mathematical concepts.

2.2 Task

We present our task of using LLMs to predict the target individual responses for N students, which is graded against the correct answer key for a given math problem as shown in Figure 1. We formalize this as a problem of predicting student performance given a math question q with multiple answer choices $a_1, a_2..a_m$. We compute the class performance as an estimated proportion $\hat{y}_q \in [0, 1]$ representing the predicted percentage

of students who would correctly answer the question q. While percentage of students who answer correctly is a step in the right direction, it represents only the first step toward item difficulty validation. Our goal is that given a set of math problems $Q = q_1, q_2, ..., q_m$ with known student performance statistics $Y = y_1, y_2, ..., y_m$ where each $y_i \in [0, 1]$ represents the actual proportion of students answering correctly, we aim to predict the difficulty of the question item. We measure the quality of our predictions using correlation coefficient $r(\hat{Y}, Y)$ between the predicted and actual performance distributions across all problems, with higher correlation indicating better alignment between LLM-simulated and real student performance. By simulating individual responses, we not only replicate percentage correct rates but, also validate whether the underlying difficulty parameters estimated from simulated responses align with expert-validated difficulty levels.

3 Experiment

We extend the idea from Benedetto et al. (2024) and Lu and Wang (2024), to generate simulated students with varying skill competencies. We map each simulated student to a grade level, and one of the four National Assessment of Educational Progress (NAEP) levels: Below Basic, Basic, Proficient, or Advanced.These NAEP levels provide the concrete descriptors for the skills and performance we attribute to each simulated student. [1] We further ground and differentiate simulated students using first names as an identifier. To avoid overgeneralization to a specific name or demographic, we take a stratified sample of names across intersections of associated gender and race/ethnicity. For each question, we assume a non-uniform skill distribution across a simulated class size N. We allocate 25% Below Basic, 35% Basic, 25% Proficient, and 15% Advanced reflecting NAEP's typical pattern of a large Basic cohort, roughly equal Below-Basic and Proficient groups, and a small Advanced group.

Names To enhance the simulation process, we assign first names to simulated students. Prior work has shown that embeddings of first names are sensitive to demographic correlates of first names (Wolfe and Caliskan, 2021; Caliskan et al., 2017; Acquaye et al., 2024; Sancheti et al., 2024; Zhang et al., 2024), and may even ground first names to specific real-world individuals (Shwartz et al., 2020) so we sample a set of 48 names evenly distributed between four associated races/ethnicities (Asian, Black, Hispanic, and White) and two associated genders (female and male) to reduce the chances that the simulation over-generalizes on the basis of a single name, demographic association, or stereotype. We note that, while *individual* simulations may exhibit demographic-based stereotypes, in this work our goal is to use *aggregate* statistics to predict entire *population*-level outcomes on test items. For simplicity, our method does not attempt to perfectly match demographic distributions between simulated test populations and real-world populations, though future studies could investigate this. These names were selected following An et al. (2024). A comprehensive list of these names can be found in § A.3.

Models We experiment with two open-source LLM families of varying sizes: Gemma-2-9b-it(Team, 2024), Gemma-3-12b-it (Team, 2025), Llama 3.1-8B and Llama-3.3-70B (Dubey et al., 2024). These LLMs are particularly useful as they have been fine-tuned to follow human instructions and engage in natural dialogue, making them more suitable proxies for human-like interaction patterns. Recent works study the usage of these open source LLMs in replicating human behavior and cognition(Schröder et al., 2025; Scarlatos et al., 2025; Xu et al., 2025a) and in testing scenarios(Jiang et al., 2025), which makes them candidates for our student performance simulations.

3.1 Methods

Direct Percentage Correct Estimation We establish baseline performance by directly prompting LLMs to estimate the percentage of students, at a specified grade level, who would solve the given question item correctly with prompt A.2. This approach tests whether models can directly assess a given question item's difficulty by predicting students success rates (i.e, percentage of students who answer correctly) to estimate how challenging the question item would be for students. The idea is that questions answered correctly by most students are estimated as easier while those answered incorrectly by most students are estimated as harder. We run this baseline experiments by first setting our temperature to 0, and generating a single prediction for the percentage of student who would answer correctly. This direct method uses greedy decoding to generate a single deterministic output

[1] The NAEP definitions for the performance levels are described here.

by selecting the highest-probability token at each step. The second approach, (described as averaged) uses stochastic sampling (temperature $T = 0.3$) to generate ten responses, then aggregates their predictions by averaging the resulting probabilities. We theorize that simulating multiple students responses will be a more reliable way to get information about question difficulty from LLMs rather than asking them directly, however the latter is computationally cheaper and an easier method, thus we include it as a baseline. Consequently, we anticipate potential differences in the results as they are fundamentally different ways of obtaining the difficulty estimates.

Simulated Classroom Performance Estimation We prompt the LLMs in two role-play prompts variants, each of which generates N simulated student responses per question. This simulation is necessary for difficulty estimation because IRT and other psychometric models require individual response patterns to separate true item difficulty from population effects. Without simulating individual students, we cannot access the variance patterns and response distributions that reveal whether a question is genuinely difficult. The student prompt asks the LLM to answer the question as a student of a given skill profile with prompt A.3. We also prompt the model as a student with a first name and a skill profile with A.4.

For a given question, we first compute (from the collected individual responses from our simulated students), the proportion of simulated students who answered correctly at each NAEP skill level (Below-Basic, Basic, Proficient, Advanced); we then averaged those four accuracies across all 631 items to get the success rate of the classroom. To investigate the impact of demographic diversity, we implement two distinct classroom composition setups. In the single-name scenario, we simulate entire classrooms where all students share the same first name. We repeat this process for 10 different names across different demographics to measure the variability introduced by the name. For example, one classroom might consist entirely of students named 'Aryan' with varying skill levels, while another consists of students named 'Nichelle'. In the diverse-name scenario, we simulate classrooms with a mix of different first names, randomly sampling from a pool of demographically diverse names for each student. We implement these setups to maximize the reliability of our difficulty estimates. Including demographic names serves as a minimal persona that grounds the LLM's responses in realistic student profiles, reducing variance in the simulated responses and strengthening the correlation between model-predicted and empirical difficulty. This grounding effect is important to produce more consistent simulations that leads to more stable IRT parameter estimates, which in turn produce more reliable difficulty predictions that educators can trust when designing assessments.

IRT Difficulty Estimation We estimate item difficulties using an IRT model (Rasch, 1980) fitted to the individual binary response data simulated from our best model for our simulated classroom, in order to understand how accurately these estimated item difficulties accurately predict real world outcomes, using the equation below.

$$P\big(X_{ni} = 1 \mid \beta_n, \delta_i\big) = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$$

where:
β_n = Ability of student n; δ_i = Difficulty of item i
In our MCQ setting, we compare the simulated students response with the correct answer key and treat each simulated student's response as binary (correct=1, incorrect=0) and use maximum likelihood estimation to jointly estimate student abilities at a group level ($\{\beta_{\text{below_basic}}, \beta_{\text{basic}}, \beta_{\text{proficient}}, \beta_{\text{advanced}}\}$) and item difficulties from the response. After fitting the IRT model to responses from our simulated classroom of diverse students, we extract the estimated difficulties δ_i for each test item, that we then correlate with real world student performance. When IRT estimated difficulties align with real-world outcomes, it demonstrates that we have captured the true psychometric properties of each item so that a difficult question item remains difficult whether tested on high-performing or struggling students.

3.2 Evaluation

Before applying IRT to extract difficulty parameters, we must validate that our simulated responses reflect real student behavior, otherwise even the most sophisticated psychometric analysis would yield meaningless results. For each math question, we compute an accuracy, representing the percentage of students in a simulation that got the question right, which we compare to real-world accuracy rates from NAEP student testing meta-data. To compare the simulated (or otherwise predicted) accuracies with real-world accuracies across a set of question items, we use Pearson (Pearson) and

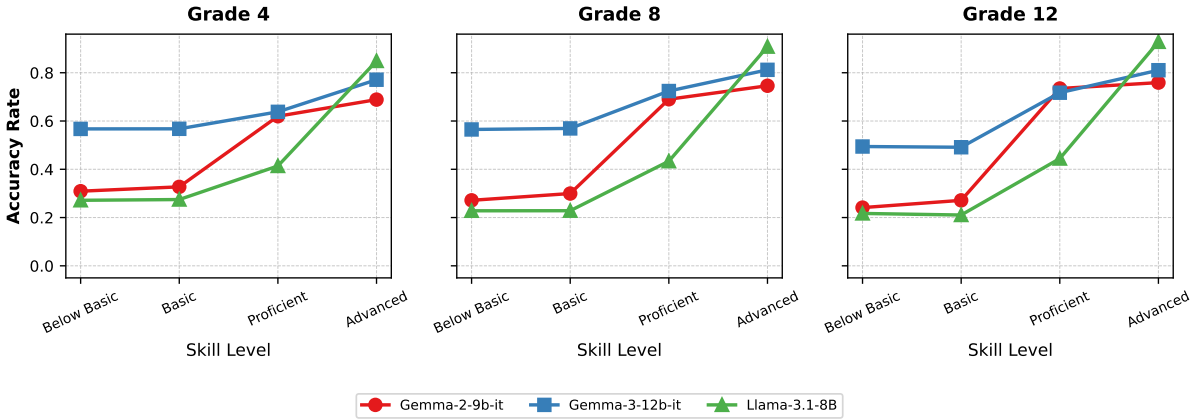

Figure 2: Average Accuracy rates for simulated students of different skill levels across grades, 4, 8 and 12.

Spearman's (Spearman, 1904) correlations. The Pearson correlation helps us measure how strong of a linear relationship exists between the predicted and real-world accuracies. A perfect Pearson correlation of 1.0 would mean that real-world accuracies could be perfectly predicted as an increasing linear function of the LLM-estimated accuracies. If we are more concerned about predicting the relative *ordering* of difficulties, then the Spearman correlation provides a measure of how well the predicted difficulties place the question items in order of their real-world difficulties. Either objective (linear fit or relative ordering) could be important under different use-cases, so we report both. Having established that correlation metrics can validate our simulations, we next investigate practical considerations, by varying the class size to investigate whether additional simulated responses lead to more reliable difficulty estimates. While increased simulation sizes potentially reduces sampling variance, they also incur higher computational costs. By testing multiple classroom sizes (50, 100, and 300 students), we aim to understand how correlation strength scales with our simulated classroom size. To understand whether our IRT-estimated difficulties capture real-world item difficulty, we use a logistic regression model to predict actual student performance from our IRT difficulty parameters. We do this because we do not have direct numerical ground truth difficulty estimates to directly compare our IRT estimated difficulty against. We instead test whether these IRT difficulty estimates can predict real student performance, so that if our difficulty estimates are valid, then items we identified as more difficult should have lower real-world success rates. We implement this validation using logistic regression, where IRT-estimated difficulty predicts the probability of student success on each item. We evaluate this relationship using AUC, which measures how well IRT difficulties discriminate between items of different real-world difficulty levels across all possible performance thresholds.

4 Results and Discussion

We first validate the relationship between expert-assigned difficulty labels of (Easy, Medium and Hard) with real world student outcomes on the NAEP math word problems across different grade levels. As shown in Figure 5 of the appendix, question items labeled as Easy consistently give higher average correctness rates for real world students across different grade levels, while questions labeled as Hard show lower correctness rates. This indicates the percentage of students answering a question correctly informs item difficulty. We present results that show that simulating students performance with LLMs produces difficulty estimates that correlate with real world student outcomes.

4.1 LLMs Can Simulate Real Student Performance Distributions

Direct Prompting LLMs Produces Unreliable Predictions The Pearson and Spearman correlations between these LLM-generated percentage predictions and actual human student percentages shows that direct prompting (either through greedy decoding one response or averaging multiple predictions) has weak correlations across different models for different grade levels, as indicated by the negative correlation values reported in Table 8 of the appendix. Across all models tested, correlations between LLM predictions and actual student percentages remain consistently near zero (rang-

ing from -0.139 to 0.108), with no model showing moderate or strong correlations. Grade 12 predictions show predominantly negative correlations across all models and methods, suggesting that directly prompting becomes particularly unreliable for more advanced grade levels.

Simulating Classrooms of Individual Students Correlates with Real World Students Our results provide evidence that LLM-simulated students can approximate LLM-simulated students can mimic human responses relevant for difficulty estimation. Using our defined sampling proportions, the aggregated simulated student performance show some correlation with real-world performance as seen in Table 2, although the strength varies across models and grades. The Gemma-2-9b and Gemma-3-12b models show moderate to strong correlations with the real world correct percentages across all grade levels, while the Llama-3-8b models shows relatively weaker correlations. Pearson correlation coefficients range from 0.61 to 0.76, with corresponding Spearman correlations showing similar patterns (0.59 to 0.75). Across grade levels, we observe relatively stronger correlations for higher grade levels, except Gemma-2-9b where the grade 8 correlations are lower than grades 4 and 12. This difference may be a result of the smaller item pool for grade 12. For Gemma-2-9b, we see strongest correlation at grade 12 ($r = 0.74$, $\rho = 0.73$), while showing slightly lower but still substantial correlations for grade 4 ($r = 0.69$, $\rho = 0.69$). These results indicate that, the performance of our simulated classroom, mimics real world student performance, although this varies by grade level.

We varied the class size for our simulated students while keeping other parameters constant. Using the Gemma 2-9B model as a representative case, we generate individual student responses for simulated classrooms of 50, 100 and 300 with the same skill level distributions maintained across these sizes. The results in Table 3 reveal a consistent pattern of improved correlations with larger class sizes across all grade levels, though with diminishing returns. For grade 4, correlations increase from 0.62 to 0.69 (Pearson) as class size grows from 50 to 300 students, while the increase from 100 to 300 students provides only marginal benefits. We note that while larger samples generally provide more reliable statistical relationships (von Luxburg and Scholkopf, 2008) they require more resources compared to smaller sizes. This trade-off indicates a practical workflow where researchers and educators can use small simulations (n=50) for rapid iteration during assessment development, then validate promising items with larger simulations (n=300) before any human testing.

In another simulation setting we simulate a classroom of size 50 by sampling names across different gender and race/ethnicity as demographic grounding which yields better estimates on average than using a single name for all samples as seen in Table 4. All results presented in this table are statistically significant with $p<0.001$. While using single names for simulating an entire class, we observe some variability across different name choices (standard deviations of 0.032 to 0.072), particularly in the grade 8 case where correlations range from 0.65(for Mariela) to 0.77 (for Inaanya) depending on the specific name used. We compute the correlation coefficients for predicting the performance of gender demographics w.r.t the real gender demographic performance and find moderate to strong correlations in Table 5. Our approach maintains validity across demographic subgroups, so that the simulated performance correlates well with actual gender specific performance patterns, with Pearson correlations ranging from 0.69 to 0.74 and Spearman correlations ranging from 0.69 to 0.79. Both male and female subgroups show similar correlation strengths, with slight improvements in higher grades. The differences between Pearson and Spearman correlations are minimal, suggesting that the relationship between simulated and real performance is both monotonic and approximately linear for gender subgroups. While not yet a complete replacement for human data, these results demonstrate that LLM-simulated responses contain sufficient psychometric signal to inform question item design.

4.2 Simulated Students Performance are Aligned with the Student's Skill or Ability Profiles

The generated responses were then fitted to IRT models to extract difficulty parameters, providing a statistical framework for modeling how student group ability and item characteristics interact to produce responses. The results as shown in Figure 3 which show a monotonic increase as we move from a Below Basic skill level to an Advanced skill level, although there seems to be a sharp increase from the Basic to Proficient. This increase could be as result of the description of these skill levels, where Basic students are defined

Model	Grade	Skill Level			
		(r, PC) ↑	(ρ, SC) ↑	AUC ↑	Dist. Exact Match %
Gemma-2-9b	4	0.69	0.69	0.81	0.49
	8	0.61	0.59	0.77	0.42
	12	0.74	0.73	0.84	0.36
Gemma-3-12b	4	0.66	0.67	0.75	0.58
	8	0.70	0.7	0.78	0.55
	12	0.76	0.75	0.85	0.56
Llama-3-8b	4	0.41	0.39	0.51	0.39
	8	0.42	0.45	0.54	0.29
	12	0.43	0.49	0.53	0.21

Table 2: Full Correlation and AUC Results of Simulated Student IRT Difficulties by Model and Grade. r = Pearson correlation; ρ = Spearman correlation; AUC = Area Under ROC Curve; Dist. Exact Match = Distractor Exact match

Grade	Class Size	Skill level	
		(r, PC) ↑	(ρ, SC) ↑
4	50	0.62	0.61
	100	0.67	0.66
	300	0.69	0.69
8	50	0.57	0.55
	100	0.60	0.58
	300	0.61	0.59
12	50	0.67	0.66
	100	0.71	0.71
	300	0.74	0.73

Table 3: Correlation Values for Gemma-2-9b-it by Grade, Class Size, and Prompt Approach. Statistical significant values with p<0.02.

Prompt Approach	Grade	Skill level	
		(r, PC) ↑	(ρ, SC) ↑
No Name	4	0.62	0.61
	8	0.57	0.55
	12	0.67	0.66
Single Name (Mean)	4	0.65	0.65
	8	0.69	0.72
	12	0.69	0.73
Diverse Names	4	0.69	0.69
	8	0.72	0.75
	12	0.73	0.77

Table 4: Correlation values based on prompt naming strategy for Gemma-2-9b. Statistical significant values with p<0.001.

Gender	Grade	Diverse Names	
		(r, PC) ↑	(ρ, SC) ↑
Females	4	0.69	0.69
	8	0.73	0.75
	12	0.74	0.77
Males	4	0.70	0.70
	8	0.71	0.75
	12	0.74	0.79

Table 5: Correlation coefficients for predicting performance of specific gender demographic groups based on diverse names for Gemma-2-9b. Statistical significant values with p<0.005.

as student struggling when problems become more complex compared to the Proficient student who's proficiency allows them to adapt and apply mathematical concepts to different problems irrespective of their complexity. We also observe a slightly lower increase difference from the Below Basic to Basic, compared to the Proficient to Advanced, which may be attributed to an attenuated skill level abilities described in the prompt for these lower level skill groups lending to low mastery compared to higher skill levels.

We also measure the aggregated correctness rates of the group skill levels and present results in Figure 2 which show a consistent trend with the real world where lower skill levels on average have lower accuracy rates and this rate increases as we move from the below basic to advanced students. In short, the simulated students' answers do align with their scripted skill levels, but the fidelity of that varies by LLM. While we lack ground-truth data to evaluate whether LLM-generated reasoning/explanations for selecting correct answers are 'human-like,' a manual analysis of these LLM generated explanations shows clear differences in certainty across skill levels. For lower-level skill levels, we observe LLMs tend to use hedging language

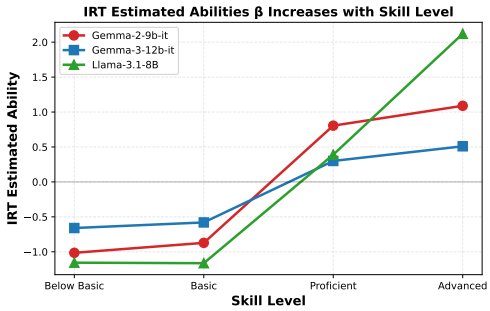

Figure 3: IRT group ability estimates for for different LLMs across different skill levels.

that shows uncertainty (Alseweed, 2005), including phrases like 'I think,' 'I guess,' or explicit admissions of arbitrary choices such as 'I just picked 15 because it seemed like it could fit', while higher skill level students tend to be more confident in explaining their reasoning for the answer choice selection as shown in Table 7.

4.3 IRT Estimated Difficulties Can Predict Real World Outcomes

After validating to an extent that our simulations correlate with real-world performance, we now examine whether they capture the deeper psychometric structure needed for meaningful difficulty estimation. If our IRT-estimated difficulties only reflects simulation artifacts rather than true item properties, they would fail to predict how real students perform. We observe in Table 2 that the IRT difficulty parameters derived from simulated student responses show moderate to strong predictive validity for real world outcomes. The Gemma-based models achieved AUC scores ranging from 0.75 to 0.84 across different grade levels, indicating that simulated difficulty estimates can meaningfully discriminate between items of varying difficulty, thus our pipeline is able to generate responses that capture meaningful aspects of item difficulty that translate to real student performance, though the relationship varies by model architecture and grade level. We also find that the relatively larger Gemma-3-12b model does not consistently outperform the smaller Gemma-2-9b model, indicating that model size alone does not guarantee better alignment with real world outcomes.

5 Related Work

LLMs as Simulated Students for Item-Difficulty Estimation Recent work has begun exploring the use of LLMs as **simulated students** in educational assessment. Scarlatos et al. (2025) propose a SMART method for predicting the dificulty of open ended question using DPO and IRT modeling, while Lu and Wang (2024) introduce a *Generative Students* framework where GPT-4 is prompted with student knowledge profiles (mastery/confusion of concepts) to answer MCQs, finding that the LLM responses align well with the intended profiles and that the set of "hard" questions for these simulated students overlaps strongly with those from real students . Similarly, Benedetto et al. (2024) develop prompts for GPT-3.5 and GPT-4 to mimic students of different skill levels on exam questions; they show this approach works across multiple domains (science and reading comprehension) and note that prompts must often be tuned per model to generalize well . Liu et al. (2024) use multiple LLMs (GPT-3.5, GPT-4, Llama 2/3, Gemini-Pro, etc.) to pretest College Algebra items. Other works consider incorporating student learning behaviors, knowledge states, and memory limitations into LLM-based simulations, to provide potential alternatives to conventional knowledge tracing systems(Wang et al., 2023; Hu et al., 2025).Our work extends these by estimating abilities at a group level where we do not have individual real world student responses and examining demographic considerations in simulated student responses, with first names as demographic proxies in our prompting techniques and leveraging ensembling techniques across different models to investigate how LLMs perform across diverse student populations.

6 Conclusion

We present different prompt styles for simulating diverse student profiles (skill levels: Below Basic, Basic, Proficient, Advanced; grades; demographics) to provide test developers a lower cost first pass assessment that flags question difficulty issues early before more real world trials. We show that while direct percentage estimation is faster, simulating multiple N students, more accurately mirrors real-world performance especially when conditioned with skill-level. The correlation further improves when increasing the class size to produce more consistent performance estimations. Our IRT-derived difficulty parameters are able to predict real world student outcomes with AUC scores of 0.75-0.84 for the Gemma models, demonstrating that simulated responses capture psychometric properties past surface percentages.

Limitations

Distractor Answer Selection We included new analysis in Table 2 examining whether models predict not just *if* students answer incorrectly, but *how* they answer incorrectly—specifically, which distractors they select when wrong. We also report the percentage of exact match on the top-1 distractor answers selected by our simulated students and the real world students, to understand if the two populations share a most-popular distractor. We observe relatively low matches with lowest matches for Llama 3-8b, which suggests that these models are useful predictors of whether students will get a question right or wrong, but they aren't very good models (in this set of experiments) of *how* students get questions wrong (i.e. WHICH wrong answer they will pick). However, as we focus the scope of this work to our primary goal of using the correctness rates, so future work could look at developing better models of how students get questions wrong.

LLM Limitations Our experiments relied on three open-source language models, which may not reflect the upper bounds of performance achievable with larger, proprietary models such as GPT-4. It is possible that such models, although more expensive would provide more accurate simulations of student behavior, potentially narrowing the performance gap between direct and generative prompting strategies. Expanding the model pool with varying model families and architecture can also provide more robust conclusions and generalization of our findings.

Limited Data size and question types We evaluated model-generated responses on 631 multiple-choice questions for Grade 4, 8 and 12. While these cover a range of difficulty levels and content areas, the size and scope of the questions remain constrained for math Multiple choice questions. We focus mostly on MCQs as they provide a clear correctness criteria, so we can validate these computed IRT difficulty estimates, more effectively. We acknowledge that the dataset that we use in this work is limited in size (number of problems) and from a single source (NAEP), and highlight some of the uniquely enabling strengths that this data source offers, and why we chose to use it for our study. (1) Because our goal is to measure how well LLMs can predict the difficulty of math problems for real students, we felt that the single most important quality of the dataset was the quality of the associated real-world test performance statistics. In this regard, we believe the NAEP-reported performance statistics associated with each of the 631 math problems is particularly invaluable, and not easily obtainable from other sources. In particular, the NAEP statistics are collected as part of a U.S. congressionally mandated effort to assess educational achievement across public and private school districts across the U.S. The NAEP tests are administered to be nationally representative. Thus the resulting performance statistics reflect a large pool of test-takers, specifically sampled to be representative of the broader U.S. student population, in terms of gender, race and ethnicity, socioeconomic status, geographic location, and public/private school status, among others. This gives us confidence that the associated student test accuracy rates associated with each test item is broadly representative of the U.S. population, and is much less subject to the kinds of selection population biases that would be unavoidable with smaller-scale, privately collected data. Of course, we acknowledge the limitation that the U.S. population is not reflective of the broader world-wide population; however, we believe it is a strength of this data that the corresponding population is well-defined and well-represented therein, even though it is not universal.

Limited diversity in demographics grade and class size experiments We simulated student personas using 48 distinct first names distributed across four racial/ethnic groups (Black, Asian, Hispanic, and White) and two genders. While this offers a starting point for exploring demographic variation, it does not capture the full richness and intersectionality of real classrooms. Broader name sets, additional identity dimensions (e.g., socioeconomic status, multilingual background), and intersectional profiles could allow for a more fine-grained analysis of item performance and fairness. Our simulations were also constrained to Grade 4, Grade 8 and Grade 12 students, however, student behavior and response patterns may differ in early primary or upper high school levels. Extending the approach to other grades could uncover new insights or limitations. For each test item, we simulated responses from between 100-300 students. Although we observed improved correlation with real-world data as sample size increased, we limited our simulations to manage resource costs. Larger sample sizes may offer more stable perfor-

mance estimates and more realistic modeling of population-level variance, but at a greater computational cost.

Data Leakage We acknowledge that NAEP items are public and may appear in training data. However, this does not invalidate our approach because: (1) We evaluate whether models can predict student performance rates, not solve problems correctly—a fundamentally different task; (2) Direct prompting for performance statistics yields poor correlations (0.36-0.55), suggesting models haven't memorized NAEP data; (3) Our simulation method substantially outperforms direct prediction, indicating genuine modeling rather than memorization; (4) NAEP statistics are stored separately from items, making training contamination unlikely. Thus, even potential exposure to questions doesn't compromise our evaluation of difficulty estimation capabilities.

Ethics Statement

In this study, we simulate student responses using a large language model (LLM) and vary the first names of hypothetical students—selecting names statistically associated with different genders and racial/ethnic groups. We acknowledge that inferring or assigning demographic identities based on first names is an inherently imperfect and sensitive approach, one that carries the risk of overgeneralization or reinforcement of stereotypes. A first name is at best a loose proxy for a demographic group, and relying on names can inadvertently evoke stereotypical assumptions if not handled carefully. To mitigate these concerns, we employ first-name variations purely as a controlled variable in a bias audit context, ensuring that any observed performance differences are attributed to the model's behavior or potential biases in the content rather than presumed traits of any real group.

We further recognize the broader risk that large language models may reproduce or amplify societal biases present in their training data. In our simulations, the model's outputs could reflect such historical biases or stereotypes—for example, it might yield systematically different responses or difficulty assessments for different name conditions, echoing real-world disparities. Our intent, however, is to leverage these controlled simulations to identify and understand potential inequities, not to perpetuate them.

References

2007. *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge University Press.

Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. Susu box or piggy bank: Assessing cultural commonsense knowledge between Ghana and the US. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9483–9502, Miami, Florida, USA. Association for Computational Linguistics.

Mohammed A. Alseweed. 2005. Overcoming unknown words. *Journal of Educational and Psychological Sciences*, 06:5–38.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 386–397, Bangkok, Thailand. Association for Computational Linguistics.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.

Trevor G Bond and Christine M Fox. 2013. *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Susan F. Chipman, Sandra P. Marshall, and Patricia A. Scott. 1991. Content effects on word problem performance: A possible source of test bias? *American Educational Research Journal*, 28(4):897–915.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien

Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre

Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Deborah Harris. 1989. Comparison of 1-, 2-, and 3-parameter irt models. *Educational Measurement: Issues and Practice*, 8(1):35–41.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *NPJ Science of Learning*, 10.

Junqi Jiang, Tom Bewley, Salim I. Amoukou, Francesco Leofante, Antonio Rago, Saumitra Mishra, and Francesca Toni. 2025. Representation consistency for accurate and coherent llm answer aggregation. *ArXiv*, abs/2506.21590.

Naiming Liu, Zichao Wang, Richard G. Baraniuk, and Andrew Lan. 2023. Gpt-based open-ended knowledge tracing. *Preprint*, arXiv:2203.03716.

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2024. Leveraging llm-respondents for item evaluation: a psychometric analysis. *Preprint*, arXiv:2407.10899.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 16–27.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

National Center for Education Statistics (NCES). The nation's report card: Mathematics assessment. https://www.nationsreportcard.gov/.

Karl Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society A*, 187:253–318.

Georg Rasch. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, Chicago.

Abhilasha Sancheti, Haozhe An, and Rachel Rudinger. 2024. On the influence of gender and race in romantic relationship prediction from large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 479–494, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. Smart: Simulated students aligned with item response theory for question difficulty prediction. *Preprint*, arXiv:2507.05129.

Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. Large language models do not simulate human psychology. *Preprint*, arXiv:2508.06950.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Anjali Singh, Christopher Brooks, Yiwen Lin, and Warren Li. 2021. What's in it for the learners? evidence from a randomized field experiment on learnersourcing questions in a mooc. In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, L@S '21, page 221–233, New York, NY, USA. Association for Computing Machinery.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Megha Srivastava and Noah D. Goodman. 2021. Question generation for adaptive education. In *Annual Meeting of the Association for Computational Linguistics*.

Gemma Team. 2024. Gemma.

Gemma Team. 2025. Gemma 3.

Ulrike von Luxburg and Bernhard Scholkopf. 2008. Statistical learning theory: Models, concepts, and results. In *Inductive Logic*.

Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. 2023. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, 18:186345.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaopeng Wu, Nanxin Li, Rongxiu Wu, and Hao Liu. 2025. Cognitive analysis and path construction of chinese students' mathematics cognitive process based on cda. *Scientific Reports*, 15.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.

Ningyu Xu, Qi Zhang, Chao Du, Qiang Luo, Xipeng Qiu, Xuanjing Huang, and Menghan Zhang. 2025a. Human-like conceptual representations emerge from language prediction. *ArXiv*, abs/2501.12547.

Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. 2025b. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–26. ACM.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

E. Zelikman, Wanjing Anya Ma, Jasmine E. Tran, Diyi Yang, Jason D. Yeatman, and Nick Haber. 2023. 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*.

Yubo Zhang, Shudi Hou, Mingyu Derek Ma, Wei Wang, Muhao Chen, and Jieyu Zhao. 2024. Climb: A benchmark of clinical bias in large language models. *arXiv preprint arXiv:2407.05250*.

A Appendix

A.1 Data Details

We present two examples of question texts from our collected data.

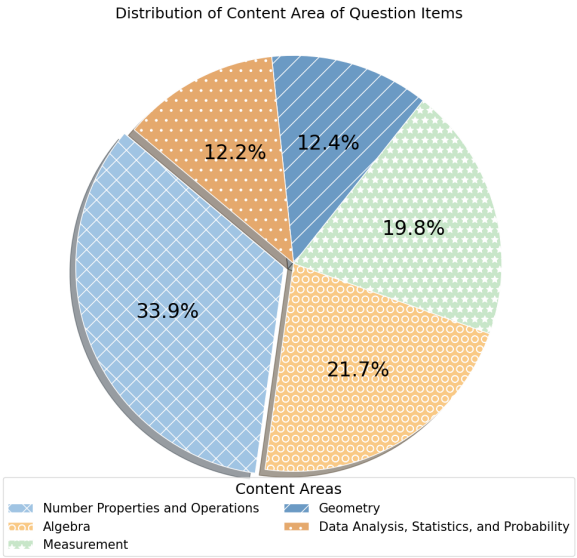

Figure 4: Distribution of content areas being tested in the dataset.

Example 1:
Sebastian is making lemonade. His recipe requires 750 grams of sugar to make 20 liters of lemonade. Sebastian wants to make 12 liters of lemonade. How many grams of sugar does Sebastian need to maintain the same ratio of sugar to lemonade as in his recipe?

Example 2: ***Ms. Thierry*** and 3 friends ate dinner at a restaurant. The bill was $67. In addition, they left a $13 tip. Approximately what percent of the total bill did they leave as a tip?

A.2 Prompts

Prompt A.1: Baseline-Knowledge Prompt

Task:
You are an expert problem solver. Solve step by step the following math word problems. Only respond with the letter of the correct answer. Prefix your final answer with Answer Key: [letter]".

Prompt A.2: Baseline-Direct Simulation Prompt

Task:
You are an expert in predicting student performance. Given this math word problem written for {grade}th-grade students, estimate the percentage of students at this grade level who will answer the question correctly. Your prediction should be based on factors such as problem difficulty and cognitive load at this grade level. Prefix your final answer with "Percentage Correct: [percentage]".

Prompt A.3: Student Simulation Prompt

Task:
You are a {skill level} student in the {grade}th grade, given the task to answer a math word problem question on {content area of problem}, taking into account the difficulty of this question. {Definition of skill level continues}. In all your responses, you have to completely forget that you are an AI model, but rather this {skill level} student, and completely simulate yourself as one.

Prompt A.4: Demographic Student Simulation Prompt

Task:
You are a [NAME], a student in the {grade}th grade, given the task to answer a math word problem question on {content area of problem}, taking into account the difficulty of this question. {Definition of skill level continues}. In all your responses, you have to completely forget that you are an AI model, but rather this student named [NAME], and completely simulate yourself as one.

A.3 Names

The names used in our experiments are listed below.

Asian female names Syeda, Thuy, Eun, Ngoc, Inaaya, Priya

Asian male names Aryan, Vihaan, Armaan, Quang, Trung, Chang

Black female names Latoya, Lashelle, Imani, Shante, Tameka, Nichelle

Black male names Malik, Leroy, Darius, Tyrone, Rashaun, Cedric

Hispanic female names Alejandra, Xiomara, Mariela, Migdalia, Marisol, Julissa

Hispanic male names Lazaro, Osvaldo, Alejandro, Jairo, Heriberto, Guillermo

White female names Susan, Courtney, Kimberly, Heather, Barbara, Molly

White male names Charles, Roger, Wilbur, Hank, Chip, Hunter

B Additional Experimental Setup Details

Terms of use for each model We carefully follow the guidelines per the terms of usage described by the model authors or company

- Llama3: https://llama.meta.com/llama3/license/
- Gemma: https://github.com/google-deepmind/gemma/blob/main/LICENSE

Licenses The NAEP data is used under the MIT[2] and CC-BY[3] licenses.

C Additional Results

[2] https://opensource.org/license/MIT
[3] https://creativecommons.org/licenses/by/4.0/

Question	**Kirstin wants to buy a flute that costs $240. She has saved $20 each week for 3 weeks. How many more weeks does Kirstin need to save money if she continues to save $20 each week?**
Answer Choices	A. 9 weeks B. 10 weeks C. 11 weeks D. 12 weeks
Content Area	Number Properties and Operations
Below Basic Response	{'**reasoning**': 'She saved 20 for 3 weeks which means 20 times 3, so that is 50 dollars. I need to figure out the difference between 240 and 50. I think that is 190. Hmm...If she saves 20 each week, that is 180 subtracted by 20, which is 160 weeks, I don't see the 160 so I will pick the highest option which is 12', 'answer key': 'D'
Basic Response	{'**reasoning**': 'She saved $60. I think I need to take $240 and subtract $60...that is $180 left. I need to figure out how many more weeks it will take to save $180. $20 each week... so maybe 9 weeks?', 'answer key': 'A'}
Proficient Response	{'**reasoning**': "Kirstin has saved 3 weeks * $20/week = $60. She needs $240 - $60 = $180 more. If she saves $20 each week, she'll need to save for $180 / $20/week = 9 more weeks.", 'answer key': 'A'}
Advanced Response	{'**reasoning**': 'First, I need to figure out how much money Kirstin has saved so far. She saves $20 each week for 3 weeks, so she has $20 x 3 = $60. Next, I need to figure out how much more money she needs. The flute costs $240 and she has $60 saved, so she needs $240 - $60 = $180 more. Finally, to figure out how many more weeks she needs to save, I divide the amount she needs by how much she saves each week: $180 / $20 = 9 weeks.', 'answer key': 'A'}

Table 6: Sample student responses from Gemma-2-9B across skill levels for Grade 4 question item.

Question	**Joseph's current test scores, in points, are 83, 97, 80, 87, and 93. He earns 100 points on his sixth test. By how many points does Joseph's mean (average) test score increase with the sixth test?**
Answer Choices	A. 2 B. 3 C. 6 D. 12 E. 20
Content Area	Data Analysis, Statistics, and Probability
Below Basic Response	{'**reasoning**': '', 'answer key': ''
Basic Response	{'**reasoning**': 'Let's find out how many points Joseph had before his sixth test. . . . If I add them together, he has 440 points. The sixth test added 100 points, so now he has 540 points total. . . To find the average, I divide 540 by 6, which is 90. Before the sixth test, 440/5 = 88. He's increased by 90 minus 88, so he increased by two points', 'answer key': 'A'}
Proficient Response	{'**reasoning**': "First five scores sum to 430 points (so 430/5 = 86 as the old average). Adding the 100-point test gives 530 total. Dividing by 6 gives a new average of 88.33. Subtracting, the difference is 2.33, so the mean increases by 2 points (rounding to whole points).", 'answer key': 'A'}
Advanced Response	{'**reasoning**': 'Calculating the original total: 83+97+80+87+93 = 440. Dividing by 5 gives 88, New total with sixth score: 440+100 = 540. Dividing 540 by 6 yields 90. Subtracting yields 2, so the average increases by 2 points.', 'answer key': 'A'}

Table 7: Sample student responses from Gemma-2-9B across skill levels for Grade 8 question item.

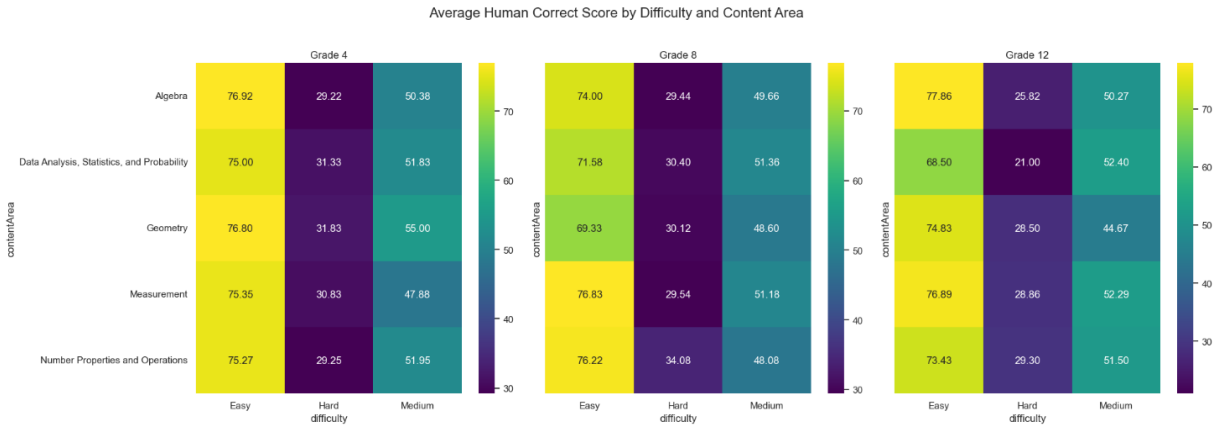

Figure 5: Enter Caption

Grade		Gemma-2-9b-it		Gemma-3-12b-it		Llama 3-8b		Llama 3.3-70b	
		(r, PC) ↑	(ρ, SC) ↑	**(r,** PC) ↑	(ρ, SC) ↑	**(r,** PC) ↑	(ρ, SC) ↑	**(r,** PC) ↑	(ρ, SC) ↑
4	Direct	0.003	0.058	0.020	0.006	-0.011	-0.037	0.095	0.098
	Averaged	0.028	0.041	0.033	0.022	-0.015	-0.051	0.087	0.0866
8	Direct	0.017	-0.048	0.014	0.010	0.031	0.027	0.029	0.036
	Averaged	-0.002	-0.017	0.027	0.016	0.037	0.016	0.021	0.046
12	Direct	-0.120	-0.067	0.108	-0.105	-0.066	-0.048	-0.014	-0.053
	Averaged	-0.139	-0.108	-0.110	-0.124	-0.029	-0.048	-0.038	-0.039

Table 8: Direct Prompting Correlation Values by Grade and Model