Take Out Your Calculators: Estimating the Real Difficulty of Question Items with LLM Student Simulations

Anonymous ACL submission

Abstract

Standardized math assessments require expensive human pilot studies to establish the difficulty of test items. We investigate the predictive value of open-source large language models (LLMs) for evaluating the difficulty of multiple-choice math questions for real-world students. We show that, while LLMs are poor direct judges of problem difficulty, simulation-based approaches with LLMs yield promising results under the right conditions. Under the proposed approach, we simulate a "classroom" of 4th, 8th, or 12th grade students by prompting the LLM to role-play students of varying proficiency levels. We use the outcomes of these simulations to fit Item Response Theory (IRT) models, comparing learned difficulty parameters for items to their real-world difficulties, as determined by item-level statistics furnished by the National Assessment of Educational Progress (NAEP). We observe correlations as high as 0.75, 0.76, and 0.82 for grades 4, 8, and 12, respectively. In our simulations, we experiment with different "classroom sizes," showing tradeoffs between computation size and accuracy. We find that role-plays with named students improves predictions (compared to student ids), and stratifying names across gender and race further improves predictions. Our results show that LLMs with relatively weaker mathematical abilities (Gemma) actually yield better real-world difficulty predictions than mathematically stronger models (Llama and Qwen), further underscoring the suitability of these models for the task.

1 Introduction

The accurate estimation of question difficulty is key for developing reliable educational assessments as difficulty is essential in determining the quality and validity of these test items (Alkhuzaey et al., 2023). Multiple-choice mathematics question items which measure mathematical ability independent of contextual knowledge, requires rigorous validation

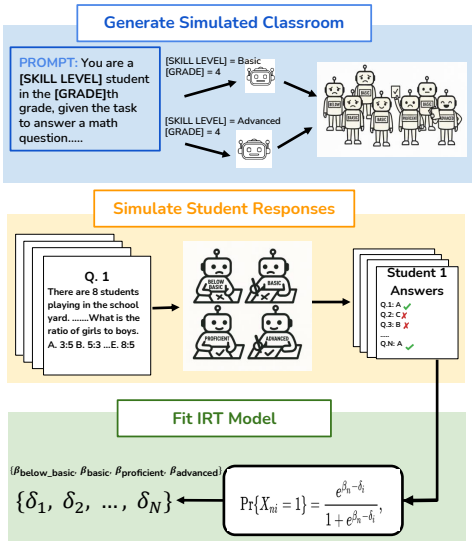


Figure 1: We simulate classrooms by prompting LLMs at different skill levels. Simulated students with varying abilities respond to question items . An IRT model is fit to the simulated responses to estimate students' ability, β and item difficulty, δ.

of difficulty to be used in assessments. Traditional approaches rely on expert cognitive analysis (Leighton and Gierl, 2007; Wu et al., 2025) or retrospective psychometric studies of student performance through pilots (Harris, 1989; Bond and Fox, 2013). However, these pilot studies are costly, time-consuming, and difficult to scale with costs reaching thousands of dollars and weeks to complete (National Academies of Sciences and Medicine, 2022).

Automated approaches like Question Difficulty Estimation from Text (QDET) offer a potential solution, but produce weak correlations with real world difficulty estimation(Scarlatos et al., 2025). Even with chain-of-thought and stronger closed-source models, direct text-based prediction fails to capture the misconception patterns underlying item difficulty (Li et al., 2025). As difficulty inherently depends on who answers, not just what is asked

(Martínez-Plumed et al., 2022), we need individual-level responses for established framework Item Response Theory (Hambleton, 1985) to estimate both difficulty and discriminability parameters, that direct estimation cannot provide. This requires costly individual response data (Zelikman et al., 2023). Even when such data exist, they are often proprietary or provide only aggregate metrics rather than the individual response IRT requires, thus a need to simulate realistic student responses.

Recent work shows that LLMs can simulate human-like responses across diverse tasks, including preliminary explorations in assessments (Lu and Wang, 2024; Xu et al., 2025b; Liu et al., 2023). We extend this direction by prompting open-source models, requiring no additional training data/fine-tuning, and validate difficulty estimates using IRT.

We simulate diverse student profiles (varying skill levels from Below Basic, Basic, Proficient, Advanced and identity attributes including student ids and demographic names) and validate these simulations against real NAEP mathematics data using IRT (Figure 1). We study: **(1)** To what extent do LLM-simulated student responses mimic real-world difficulty **(2)** How do prompt design, identity attributes and model selection affect simulated vs real student performance alignment? We find correlations as high as 0.75, 0.76, and 0.82 for grades 4, 8, and 12, respectively, with LLMs with relatively weaker mathematical abilities and an ensemble of top performing models falling within acceptable range by standard reliability benchmarks (Cicchetti, 1994). Diverse identity improves correlations over no-identifier baselines and IRT-estimated difficulties achieve AUC scores of 0.78–0.90, suggesting that simulations can discriminate between items that real students find easy versus hard. LLMs with relatively weaker mathematical abilities outperform highly-proficient models, suggesting that these proficient models struggle to reproduce realistic error patterns. These findings show open source LLM simulations can offer a practical low-cost screening tool, with the counterintuitive finding that LLMs with relatively weaker mathematical abilities often better simulate struggling students.

2 Related Work

Recent work show LLMs can act as "silicon" subjects, reproducing human behavioral patterns across diverse tasks (Xie et al., 2024; Argyle et al., 2023; Dillion et al., 2023; Manning et al., 2024; Yang et al., 2024). Scarlatos et al. (2025) propose SMART for difficulty prediction of open ended question using Direct Preference Optimization (DPO) and IRT, while Lu and Wang (2024) introduce a *Generative Students* framework where GPT-4 is prompted with student knowledge profiles to answer Multiple Choice Questions (MCQs), finding that the LLM responses align with the intended profiles and that the set of "hard" questions align with real students' difficulty. Similarly, Benedetto et al. (2024) develop skill-level prompts for GPT-3.5/4 across domains, noting model-specific tuning requirements. Liu et al. (2024) use multiple LLMs (GPT-3.5/4, 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 simulations, to provide potential alternatives to conventional knowledge tracing systems (Wang et al., 2023; Hu et al., 2025). Our work extends these with open source models, by estimating abilities at a group level where we do not have individual real world student responses, examining student ids and first names as demographic identifiers in our prompting techniques and leveraging ensembling techniques across models.

3 Preliminaries

3.1 Data

We collect 631 maths MCQs 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 questions provide real student 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 USA. 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 demographics. The items cover a diverse range of mathematical concepts. Figure 3 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. Table 6 in the appendix summarizes the distribution of question items by grade level and difficulty.

3.2 Task

Our task is to use LLMs to predict classroom performance by simulating how N students with varying skill levels would answer a given math question. Each simulated response is graded as correct or incorrect against the answer key as shown in Figure 1. Given a math question, q with multiple answer choices $\{a_1, a_2..a_m\}$, we generate N individual student responses. We then aggregate these responses to compute the simulated classroom success rate as an estimated proportion $\hat{y}_q \in [0, 1]$ representing the predicted percentage of simulated students who would correctly answer the question q. To evaluate our approach, we test whether simulated performance aligns with real-world data. Given a set of $Q = \{q_1, q_2, ..., q_m\}$ with student performance statistics $Y = \{y_1, y_2, ..., y_m\}$ (where each $y_i \in [0, 1]$ represents the actual percentage of students answering q_i correctly), we compute the correlation $r(\hat{Y}, Y)$, between our simulated success rates $\hat{Y}$ and real world success rates, Y. In addition to percentage correct rates, we ground our approach using Item Response Theory (IRT), a well established psychometric framework that estimates item difficulties (δ) and student abilities (β) by modeling response probabilities. IRT requires individual student responses, to separate true item characteristics from population-level effects. We fit our IRT model to our simulated N students responses and test whether the estimated difficulty parameters predict real world outcomes, showing psychometric validity beyond simple correlations.

4 Experiment

We extend 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 also simulate students using student ids (A.4) and first names to test whether individuating students further improves simulation quality. To avoid overgeneralization to a specific demographic, we take a stratified sample of names across intersections of associated gender and race/ethnicity. For each question, we generate responses for a simulated classroom of size N with a non-uniform skill distribution: 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 investigate whether adding identity markers improves simulation, we assign first names to simulated students. Prior work shows name embeddings are sensitive to demographic correlates of first names (Caliskan et al., 2017; Wolfe and Caliskan, 2021; Zhang et al., 2024). We sample 48 names evenly distributed between four associated races (Asian, Black, Hispanic, and White) and genders (female and male) to reduce chances of overgeneralization of the simulation on a single name, demographic association, or stereotype. 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, though future work can look at distinguishing this stereotype activation. We emphasize that we do not use these names to make predictions about specific people/groups but report empirical observations of using these names. These names were selected following An et al. (2024). A comprehensive list of these names is in § A.3.

	G2		G3			L3		Q2.5		Q3
	9B	27B	4B	12B	27B	8B	70B	14B	32B	4B
Acc. ↑	0.72	0.74	0.56	0.78	0.83	0.81	0.92	0.92	**0.93**	**0.93**

Table 1: Model Accuracy Across All Questions. G2/G3 = Gemma-2/3, L3 = Llama-3, Q2.5/Q3 = Qwen-2.5/3.

Models We experiment with ten open-source LLM from different model families and generations of varying sizes: Gemma-2 (9b and 27b) (Team et al., 2024), Gemma-3 (4b, 12b, 27b) (Team et al., 2025), Llama 3.1-8B and Llama-3.3-70B (Grattafiori et al., 2024), Qwen2.5 (14B and 32B) (Yang et al., 2025b) and Qwen3-4B(Yang et al., 2025a). These models show moderate to good math reasoning with accuracy scores (Table 1) on the NAEP questions, as well as abilities in role-playing and following instructions. Recent works study open source LLMs in replicating human behavior,

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

cognition and testing settings (Schröder et al., 2025; Scarlatos et al., 2025; Xu et al., 2025a; Jiang et al., 2025) which makes them candidates for our student simulations. We **note** that all selected models are instruction-tuned and used according to their intended purpose. However, we recognize that a model's ability to simulate different student performance (such as an advanced student who answers more correctly versus a below-basic student) will be inherently influenced by whether the model can actually solve the question correctly on its own, as shown in Table 1.

4.1 Methods

Direct Question Difficulty Estimation from Text Recent Question Difficulty Estimation from Text (QDET) (Benedetto, 2023) approaches use embeddings or transformers to directly estimate the difficulty of question items. We implemented Word2Vec-based difficulty estimation, which encodes question texts as embeddings and uses them as input to a Random Forest regression model to output continuous difficulty values. For transformer baselines, we extract [CLS] token embeddings from pre-trained BERT-base-uncased and DistilBERT-base-uncased and run inference on question text concatenated with the correct answer. This setting enables comparison with LLM prompting, where correlations are computed over all 631 items without task-specific training.

Direct Percentage Correct Estimation (DPCE) We also 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 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 by first setting our (temp. $T = 0$) and using greedy decoding to generate a single deterministic single prediction for the percentage of student who would answer correctly for each question. The second averaged approach, uses stochastic sampling (temp. $T = 0.3$) to generate ten responses per question, then aggregates their predictions by averaging the resulting percentages. We theorize that simulating multiple students responses is a more reliable way to get information about question difficulty from LLMs rather than asking them directly (Scarlatos et al., 2025), as difficulty inherently depends on who answers, not just what is asked, however the latter is computationally cheaper, thus we include it as a baseline.

Simulated Classroom Performance Estimation We prompt the LLMs in a student role-play prompt approach to generate N simulated student responses. Without simulating individual students, we cannot access the variance patterns and response distributions that reveal whether a question is genuinely difficult. A single prediction (e.g., "60% of students will answer correctly") from DCPE cannot provide the response variance patterns necessary for difficulty estimation using established psychometric IRT framework.The prompt asks the LLM to answer the question as a student of a given skill profile with Prompt A.3. For each question, we prompt LLMs to generate individual student responses using role-play prompts that specify grade level and NAEP skill level (Below Basic, Basic, Proficient, Advanced). Each simulated response is scored as correct or incorrect against the answer key. We then compute the classroom success rate: the proportion of all simulated students who answered correctly, aggregated across the skill distribution (25% Below Basic, 35% Basic, 25% Proficient, 15% Advanced). To investigate whether student identity grounding affects difficulty estimation, we vary the identifier type across four conditions: (1) No Identifier (2) Student IDs, where each student receives a randomly generated alphanumeric identifier (e.g., STU000142); (3) Single Name, where all students in a classroom share the same first name but differ in skill level. For example, one classroom might consist entirely of students named 'Aryan' with varying skill levels, while another consists of students named 'Nichelle'; and (4) Diverse Names, where students have different first names sampled across demographic groups (gender and race). This setup tests whether diverse student identities in simulated classrooms improves correlation with real-world performance using prompts A.4 and A.5. Including student ids and demographic names serves as a minimal persona cue to empirically improve correlation, though the underlying mechanism remains uncertain. This improvement could reflect increased prompt specificity, or other factors requiring further investigation. This 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 rely on when designing assessments. We investigate practical considerations, by varying the class size to investigate whether additional simulated responses lead to more reliable difficulty estimates. While increased simulation sizes reduces sampling variance, they also incur higher computational costs. By testing multiple classroom sizes, we aim to understand how correlation strength scales with our class size.

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. This 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 compare with real world difficulties. 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.[2]

4.2 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.

[2]Appendix B.1 summarizes all experiments conducted for this work.

We compute correlation between simulated and real-world accuracy rates using Pearson (linear relationship) and Spearman (rank ordering). To validate IRT-estimated difficulties, we use regression to predict difficulty class, evaluating with AUC to measure how well difficulties discriminate between easy and hard items. We do this as we do not have direct numerical ground truth difficulty estimates to directly compare our estimated difficulty against.

5 Results and Discussion

We 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. Figure 4 shows *'Easy'* questions consistently give higher average correctness rates for the real world across different grade levels, while *'Hard'* questions show lower correctness rates. This indicates the percentage of students answering a question correctly informs item difficulty. We present baseline results from direct difficulty estimation and simulated students results that show simulating students produces difficulty estimates that correlate better with the real world.

5.1 Direct Difficulty Estimation Produces Unreliable Predictions

Question Difficulty Estimation from Text (QDET) Table 8 shows Word2Vec has moderate performance (ρ=0.30-0.36 at Grades 4 and 12) but weakest at Grade 8 (ρ=-0.04). BERT and DistilBERT have relatively weaker correlations (ρ=0.01-0.21), suggesting that these text embeddings provide limited signal for our domain-specific difficulty prediction. These results show that question and answer text complexity is insufficient for accurate difficulty prediction in our NAEP context.

Direct Prompting LLMs Correlations between these LLM-generated percentage predictions and real human student percentages shows that direct prompting has weak correlations across different models for different grade levels, as indicated by the negative correlation values reported in Table 12 and 13 of the appendix. Across all models tested, correlations remain consistently near zero (ranging from -0.139 to 0.137), with no model showing moderate or strong correlations. Grade 12 predictions show predominantly negative correlations, suggesting that directly prompting becomes particularly unreliable for more advanced grades.

Model	(r, PC) ↑			(ρ, SC) ↑			AUC ↑		
	G4	G8	G12	G4	G8	G12	G4	G8	G12
Gemma-2-9b	**0.69**	0.61	0.74	**0.69**	0.59	0.73	0.84	0.78	0.84
Gemma-2-27b	**0.69**	**0.68**	**0.78**	**0.68**	**0.67**	**0.78**	**0.85**	**0.83**	**0.9**
Gemma-3-27b	**0.70**	**0.72**	0.75	**0.74**	**0.72**	0.75	**0.88**	**0.85**	**0.89**
Llama-3-8b	0.42	0.44	0.44	0.40	0.46	0.48	0.67	0.72	0.75
Qwen2.5-32B	0.60	0.62	0.51	0.64	0.65	0.50	0.82	0.81	0.76
Qwen3-4b	0.60	0.58	0.62	0.64	0.63	0.62	0.79	0.80	0.83

Table 2: Selected Models Correlation and AUC Results of Simulated Students. All correlations are statistically significant at $p < 0.01$. r = Pearson correlation; ρ = Spearman correlation; AUC = Area Under Curve. See full results for other models in Table 10

5.2 LLMs Can Simulate Real Student Performance Distributions

Simulating Classrooms of Individual Students Correlates with Real World Students Our results provide evidence that LLM-simulated students can mimic response selections relevant for difficulty estimation to an extent. Using a class size of 300 with sampling proportion (75 Below basic, 105 Basic, 75 proficient and 45 advanced students), the aggregated simulated student performance shows some correlation with real-world performance as seen in Table 2, although the strength varies substantially across models and grades. Correlation coefficients range from 0.42 to 0.78. We find the Gemma models have the relative strongest correlations, with Gemma-2-27b achieving the highest correlations at grade 12 (r = 0.78, ρ = 0.78), while the Llama models show weaker correlations, with Llama-3-8b ranging from 0.42 to 0.44.

Models That Struggle with Problems Better Simulate Student Performance A key finding is that larger and higher predictive accuracy models do not necessarily produce better simulations of student performance. Within the Gemma-2 family, the 9B parameter model achieves almost comparable correlations to the 27B model at grade 4 (r = 0.69 vs. 0.69), despite having fewer parameters. We also find, Qwen2.5-32B (one of the largest and highest predictive accuracy models tested), produces only moderate correlations (r = 0.51-0.62) suggesting that model scale alone is insufficient for simulation fidelity. Similarly, Llama-3-3-70b shows strong baseline performance (92% accuracy) but produced relatively poor simulation correlations (Table 10). This suggests that models that are adept at solving mathematical problems may struggle to authentically simulate the response patterns of struggling students. Models that find the problems more challenging may better capture the distribution of student difficulties. Across grades, we generally observe stronger correlations for higher grades, though this pattern is not universal. Most models show improvement from grade 4 to grade 12, with notable exceptions: Gemma-2-9b exhibits lower correlations at grade 8 (r = 0.61) compared to grades 4 and 12, while Qwen2.5-32B shows worse correlations at grade 12. The Gemma family shows the most consistent performance across grade levels, with Gemma-3-27b maintaining correlations above r = 0.70 and ρ=0.72 (Spearman) across all grades. We also visualize the correlation patterns in scatter plots shown in Figure 6 which show that Gemma models with stronger correlations have tighter clustering around the diagonal of perfect agreement, with the select Gemma models showing relatively clearer linear trends than Llama.

Approach	(r, PC) ↑			(ρ, SC) ↑		
	G4	G8	G12	G4	G8	G12
Averaged	0.73	0.72	0.74	0.72	0.70	0.74
Weighted	**0.75**	**0.76**	**0.82**	**0.75**	**0.75**	**0.82**

Table 3: Ensemble Model Performance Across Grades

Ensembling Models Improves Correlation Strength We ensemble models in Table 3 in an average approach (all 10 models percentage correct value is averaged for each item) and in a the weighted (equal weights) ensemble of top-performing models (Gemma-2-9B, Gemma-2-27B, Gemma-3-27B). We see that averaging all model predictions gives ρ=0.70-0.74, while our weighted ensemble of top-performing models achieves a relatively stronger correlation of ρ=0.82 at Grade 12.

Performance Varies by Content Area Content-area analysis (Table 14) show simulations excel

at Measurement (ρ=0.78-0.87) but struggle with Algebra (ρ=0.45-0.64).

Class Size	(r, PC) ↑			(ρ, SC) ↑		
	G4	**G8**	**G12**	**G4**	**G8**	**G12**
50	0.62	0.57	0.67	0.61	0.55	0.66
100	0.67	0.60	0.71	0.66	0.58	0.71
300	**0.69**	**0.61**	**0.74**	**0.69**	**0.59**	**0.73**

Table 4: Correlation values for Gemma-2-9b-it by class size across grades. Larger class sizes consistently improve correlation with real student skill levels. All values statistically significant with p<0.01.

Larger Class Sizes Improve Reliability We varied the class size for our simulated students while keeping other parameters constant. Using the Gemma 2-9B model for its speed of runs and smaller parameter size, 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 4 show a pattern of improved correlations with larger class sizes across all grade levels. For grade 4, correlations increase from (r = 0.62 to 0.69) as class size grows(50 to 300), while the increase from 100 to 300 students provides only marginal benefits. We note that while larger samples provide more reliable statistical relationships (Luxburg and Scholkopf, 2008) they require more resources compared to smaller sizes. This trade-off indicates a practical workflow where researchers can use small simulations (n=50) for rapid iteration during assessment development, and validate promising items with larger simulations (n=300) before any human testing.

Diverse Names Identifier Improves Correlation Performance We simulate a classroom of 50 students using Gemma-2-9b with different identifier strategies: no identifier, student IDs, single homogeneous names (all students share one name), and diverse names sampled across gender and race. We find that providing student identifiers improves performance over no-identifier baselines, with student IDs (Pearson: 0.60-0.72, Spearman: 0.58-0.72) outperforming no identifier (Pearson: 0.57-0.67, Spearman: 0.55-0.66) across grades(Table 5). For single homogeneous names (where all 50 students share the same name but have different skill levels), we observe some variance across individual name choices (SD=0.059-0.090), with Grade 12 correlations ranging from 0.48 (Eun) to 0.77 (Inaanya) for Pearson and 0.53 (Chip) to 0.76 (Tameka) for Spearman. Aggregating this single name choice by race shows Hispanic names have highest average performance (r: 0.64, ρ: 0.62), followed by White (r:0.63, ρ:0.61), Black (r:0.63, ρ:0.60), and Asian names (r:0.60, ρ:0.58). Notably, aggregating all single homogeneous names perform comparable to no-identifier baselines, suggesting that providing a uniform name across all students offers minimal benefit over no identifier (Table 5). However, diverse names consistently achieve the strongest correlations. Even when comparing classrooms with diverse names within single racial groups against classrooms with diverse names across all races, the fully diverse condition outperforms all race-specific diverse conditions across grades (Table 11), showing that cross-racial demographic diversity improved model's correlation. We emphasize that these empirical results examine aggregate patterns across demographics; individual name results are never used to make claims about specific people, and all analyses report only aggregate statistics to characterize systematic model behavior rather than individual-level outcomes.

Prompt	(r, PC) ↑			(ρ, SC) ↑		
	G4	**G8**	**G12**	**G4**	**G8**	**G12**
No identifier	0.62	0.57	0.67	0.61	0.55	0.66
Std. IDs	0.66	0.60	0.72	0.64	0.58	0.72
Single	0.61	0.59	0.68	0.60	0.57	0.65
Diverse	**0.72**	**0.66**	**0.74**	**0.71**	**0.64**	**0.74**

Table 5: Correlations on prompt approaches for Gemma-2-9b across grades. Prompt approaches: No Name (no student identifier), Student IDs, Single Name, Diverse Names. All values statistically significant with p<0.01.

We also observe simulated performance correlates well with real world gender performance patterns (r: 0.69 to 0.74 and ρ: 0.69 to 0.79) in Table 9 and show similar correlation strengths. 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. While not yet a complete replacement for human data, these results show that simulated responses have psychometric signal to inform item design.

5.3 Simulated Students Performance are Aligned with the Student's Skill Profiles

The generated responses fitted to IRT models to extract difficulty parameters, provide a statistical framework for modeling how student group ability and item characteristics interact to produce responses. The results in Figure 2 show a monotonic

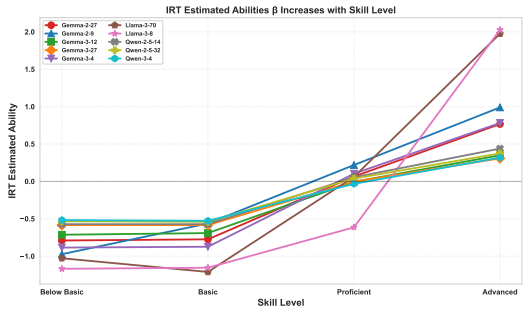


Figure 2: IRT group ability estimates for for different LLMs across different skill levels.

increase from a Below Basic skill level to an Advanced skill level, although there seems to be a more sharper increase from the Basic to Proficient. This increase could be as result of the description of these skill levels, where Basic students are defined 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 observe a slightly lower increase 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 skills lending to low mastery compared to higher skills.

We also measure the aggregated correctness rates of the group skill levels and present results in Figure 5 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 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 proficient students are more confident in their reasoning for the answer selected (Tables 16 to 18).

5.4 IRT Estimated Difficulties Can Predict Real World Outcomes

We examine if simulations 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 see in Table 2 that the IRT difficulty parameters derived from simulated student responses show moderate to strong predictive validity for real world outcomes. AUC specifically measures the models' understanding of relative difficulty (which items are harder than others). The Gemma models achieved AUC scores ranging from 0.78 to 0.90 across different grade levels, indicating that simulated difficulty estimates can meaningfully discriminate between items that real students find easy versus hard. 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 and grade. We observe that model size does not guarantee better difficulty discrimination ability.

6 Conclusion

We show that open-source LLMs can simulate students to estimate math question difficulty, achieving correlations as high as 0.75, 0.76, and 0.82 for grades 4, 8, and 12, respectively, outperforming baseline text-based and direct prompting approaches. Our simulations differentiate student skill levels (Below Basic, Basic, Proficient, Advanced) in ways that align with psychometric expectations for predicting real student outcomes. We find that diverse names improve correlations over no names or homogeneous names, though the underlying mechanism requires further investigation. Content-area analysis reveals simulations excel at Measurement problems where misconceptions follow predictable patterns, but struggle with Algebra where diverse reasoning strategies complicate simulation. Counterintuitively, LLMs with relatively weaker mathematical abilities produce better student simulations than higher mathematical proficient models, suggesting models should experience difficulty to authentically simulate struggling students. While these simulated students' computational costs are higher than direct methods (4-48 GPU-hours), these produces psychometrically valid difficulty estimates suitable for preliminary item screening before expensive human piloting.

Limitations

Distractor Answer Selection We included new analysis in Figure 7 and Table 15 examining whether simulated students select the same incorrect answers (distractors) as real students. Figure 7 reports 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. Table 15 measures the distractor match rate, the percentage of items where the most popular incorrect answer among simulated students matches the most popular incorrect answer among real students (random chance: 25% for 4-choice MCQ and 20% for 5-choice MCQ). We observe relatively low matches with all models, 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). We see a similar trend with the content area correlation results which shows the real world distractor answer selection has more distractor alignment on Measurement problems, where misconceptions follow predictable procedural patterns while, misconception matches are weakest on Algebra, reflecting greater diversity in algebraic reasoning errors. 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 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. We deliberately focus on open-source models to enable reproducibility and cost-effective deployment. While proprietary models may achieve different results, our goal is demonstrating practical screening tools accessible to educational practitioners without API costs.

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. We recognize that our use of demographic name proxies may activate stereotypes encoded in model training data. While our aggregate results show diverse names improve correlation, we cannot determine whether this reflects: (1) stereotype-based performance simulation (models enacting biased assumptions about name-associated demographics), (2) increased prompt specificity providing better grounding, or (3) other mechanisms. Individual name variance (SD=0.059-0.090) could suggest potential stereotype activation, though we emphasize all analyses report only aggregate statistics across demographic categories, never making individual-level predictions. Future work can investigate whether observed performance patterns reflect genuine diversity benefits or problematic stereotype reinforcement.

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 performance 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, 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 over-generalization 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 emphasize that these are empirical observations and that they correlate with improved alignment in our experiments.

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

Samah Alkhuzaey, Floriana Grasso, Terry R. Payne, and Valentina A. M. Tamma. 2023. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, pages 1–53.

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. 2023. A quantitative study of nlp approaches to question difficulty estimation. In *International Conference on Artificial Intelligence in Education*.

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.

Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ronald K. Hambleton. 1985. *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishers, Boston.

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.

Jacqueline P. Leighton and Mark J. Gierl. 2007. Cognitive diagnostic assessment for education: Theory and applications.

Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W Lissitz. 2025. Item difficulty modeling using fine-tuned small and large language models. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 48–55, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

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*, L@S '24, page 16–27, New York, NY, USA. Association for Computing Machinery.

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

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.

Fernando Martínez-Plumed, David Castellano Falcón, Carlos Monserrat Aranda, and José Hernández-Orallo. 2022. When ai difficulty is easy: The explanatory power of predicting irt difficulty. In *AAAI Conference on Artificial Intelligence*.

Engineering National Academies of Sciences and Medicine. 2022. *A Pragmatic Future for NAEP: Containing Costs and Updating Technologies*. The National Academies Press, Washington, DC.

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

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

Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. SMART: Simulated students aligned with item response theory for question difficulty prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25082–25105, Suzhou, China. Association for Computational Linguistics.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, et al. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

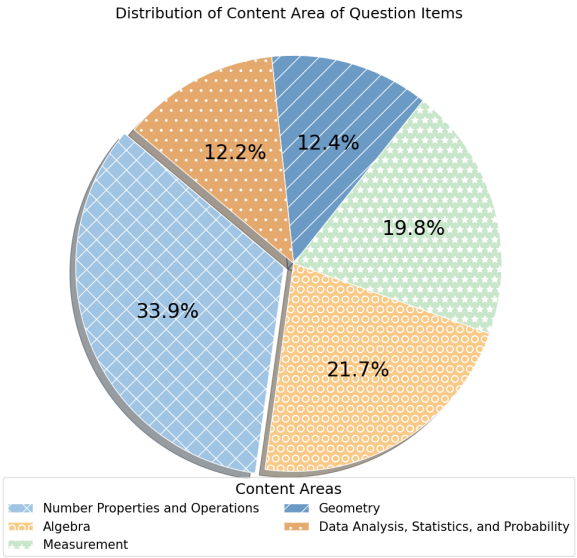


Figure 3: Distribution of content areas being tested in the dataset.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, et al. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

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?

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 6: Breakdown of question difficulty by grade.

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 [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 [NAME], and completely simulate yourself as one.

Prompt A.5: Student ID Simulation Prompt

Task:
You are [STDID], 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 [STDID], and completely simulate yourself as one.

A.3 Names

The names used in our experiments are listed.

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

A.4 Student IDs Used

We generated alphanumeric identifiers following the format `STU` followed by a 6-digit zero-padded number (e.g., `STU000001`, `STU600298`). IDs were randomly generated to avoid any systematic patterns, with randomly sampled from 000000-99999 and prepended with `STU` to create the identifier.

B Additional Experiment

B.1 Experimental Setup Details

Experimental Design Summary We formalize our simulation as a function with the parameters:

$$\text{Correlation} = f(\mathcal{M}, \mathcal{S}, \mathcal{I}, N, \mathcal{E}) \quad (1)$$

where:

- $\mathcal{M}$: **Model selection** – which LLM(s) is used
- $\mathcal{S}$: **Skill distribution** – proportion of Below Basic, Basic, Proficient, Advanced students
- $\mathcal{I}$: **Identifier strategy** – how students are identified (No Identifier, Student IDs, Single Name, Diverse Names)
- N: **Classroom size** – number of simulated students per question
- $\mathcal{E}$: **Estimation method** – how difficulty is computed (Direct prompt, Simulated responses + IRT)

Table 7 summarizes all experimental configurations and their corresponding results.

Experiment	Purpose	Key Parameters	Models
Baseline: QDET	Establish text-based difficulty estimation baseline	Method: Word2Vec, BERT, DistilBERT	BERT, DistilBERT
Baseline: Direct Prompting	Establish direct LLM prompting baseline	Method: Greedy & averaged predictions	All 10 models
Core Simulation	Evaluate simulation fidelity across models	$\mathcal{M}$: 10 models; $\mathcal{I}$: No Name; N: 300	All 10 models
Model Ensembling	Test if combining models improves correlation	$\mathcal{E}$: Averaged vs. Weighted ensemble	Averaged (all 10), Weighted (top 3)
Class Size Variation	Determine optimal classroom size	N: 50, 100, 300 students	Gemma-2-9b
Identifier Strategy	Compare naming/ID approaches	$\mathcal{I}$: None, IDs, Single Name, Diverse Names	Gemma-2-9b
Within-Category Diversity	Test racial homogeneity vs. diversity	$\mathcal{I}$: Single-race diverse vs. all-race diverse	Gemma-2-9b
Gender Correlation	Validate against gender-specific performance	$\mathcal{I}$: Diverse Names (gender-specific analysis)	Gemma-2-9b
Content Area Analysis	Identify domain-specific strengths/weaknesses	Stratify by: Algebra, Geometry, Measurement, etc.	All 10 models
Skill Level Alignment	Verify IRT ability estimates match skill levels trend expectations	$\mathcal{S}$: NAEP 4-level distribution	All 10 models

Table 7: Summary of experimental configurations. All simulations use NAEP skill distribution (25% Below Basic, 35% Basic, 25% Proficient, 15% Advanced) unless noted otherwise.

Computational Cost QDET baselines require less than 1 minute for transformer inference across all 631 items on A6000. While our simulation approach using open-source models requires approximately 4-5 GPU-hours on A6000 GPUs for models in the 8-12B parameter range to generate 300 simulated student responses across all 631 items. Larger models (e.g., 70B parameter) would require proportionally more time, though we found the correlation strengths are not comparable with model scale. For example the Gemma-2-9B needs 4.34 GPU-hours on 1 A 6000, for all 631 items for 300 students and Llama-3.3-70B requires up to 48 hours with 4 A6000s for 300 student simulation.

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

- Gemma2: https://github.com/google-deepmind/gemma/blob/main/LICENSE
- Gemma-3: https://ai.google.dev/gemma/terms
- Llama3: https://llama.meta.com/llama3/license/
- Qwen2.5/3: https://github.com/QwenLM/Qwen/blob/main/LICENSE

Licenses The NAEP data is used under the MIT[3] and CC-BY[4] licenses.

B.2 Additional Results

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

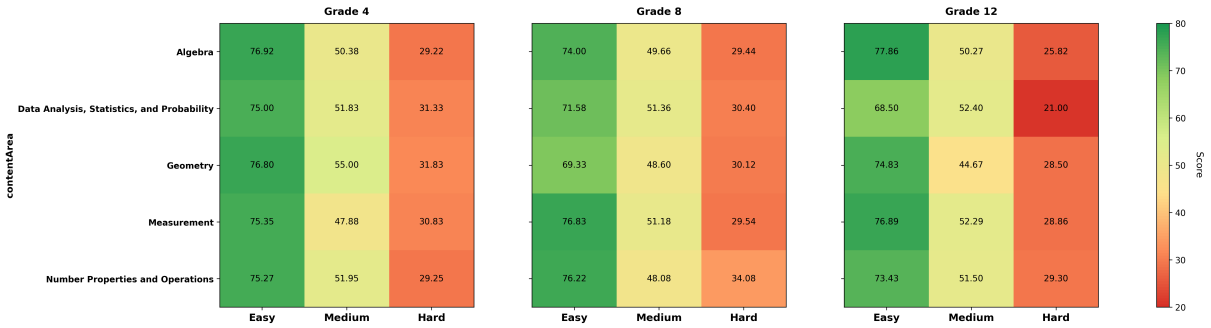


Figure 4: Average Percentage Correct Score by NAEP Assigned Difficulty at a grade and content area breakdown.

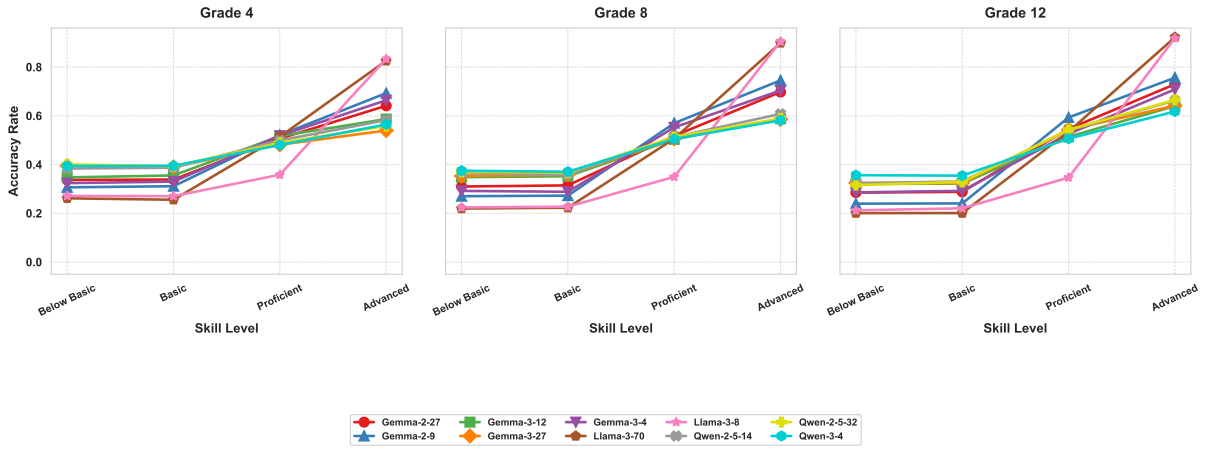
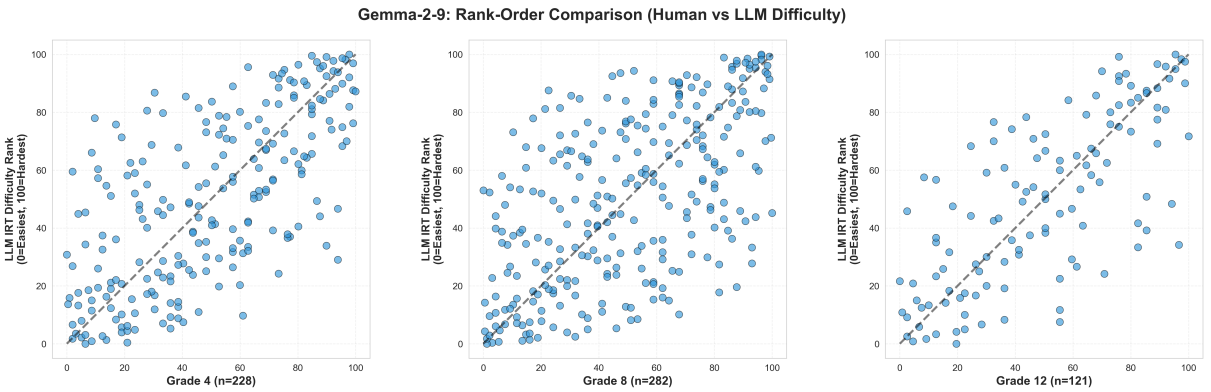


Figure 5: Average Accuracy rates for simulated students of different skill levels across grades, 4, 8 and 12.

Figure 6: Complete rank-order difficulty scatter plots for select models *(Gemma-2-9, Gemma-3-27, Llama-3-8)* across three grade levels. Each panel shows the rank-ordered difficulty of 631 test items (0=easiest, 100=hardest) for human students (x-axis) versus LLM simulations (y-axis). Tighter clustering around the diagonal indicates better preservation of relative item difficulty. These visualizations confirm that correlation strength reliably predicts visual clustering quality, with r>0.65 indicating suitable performance for preliminary item screening.



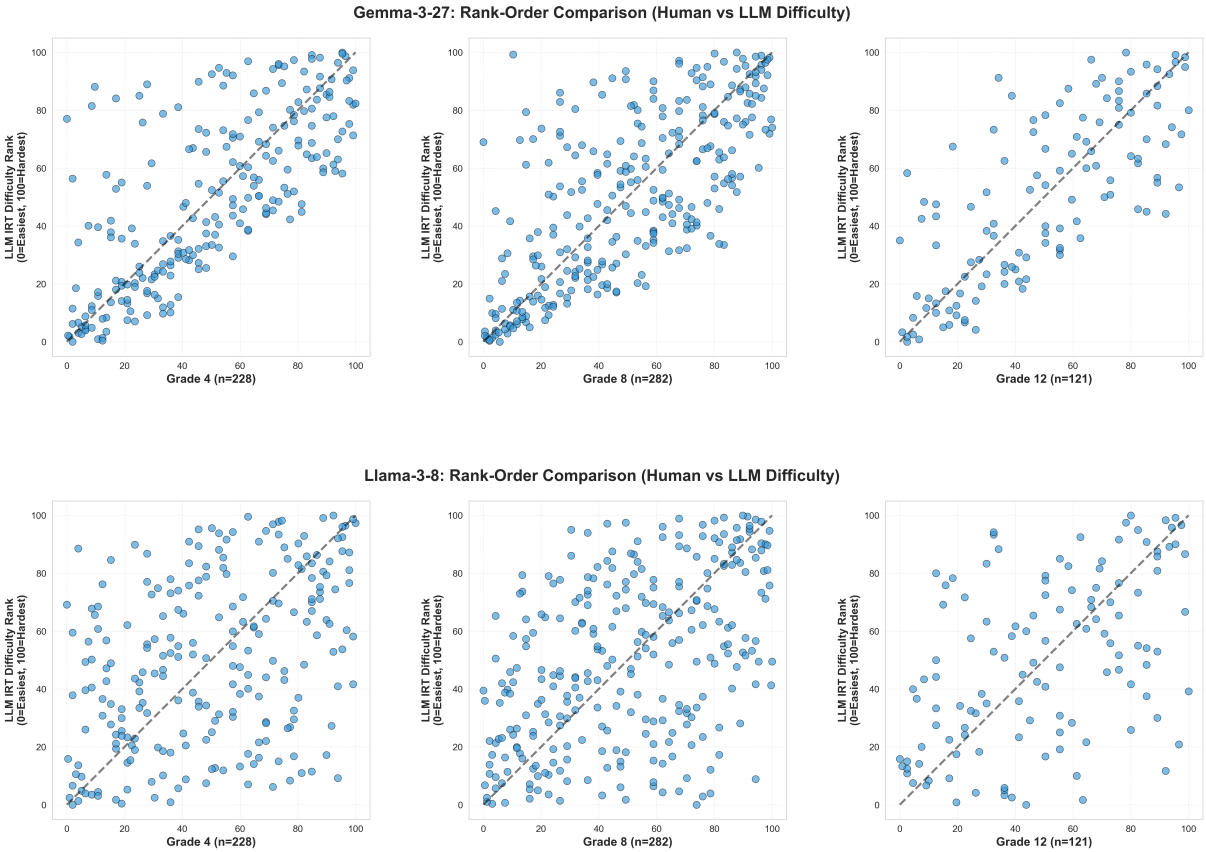


Figure 7: Full report of top distractors that matches for the real-world and llm response. The percentage shows the exact match on the top-1 distractor answers selected by our simulated students and the real world students.

Approach	Grade 4	Grade 8	Grade 12
Word2Vec + RF	0.36	-0.04	0.30
BERT	0.05	0.03	0.21
DistillBERT	0.04	0.01	0.17

Table 8: Comparison of QDET performance across different grade levels.

Gender	(r, PC) ↑			(ρ, SC) ↑		
	G4	G8	G12	G4	G8	G12
Females	0.69	**0.73**	0.74	0.69	**0.75**	0.77
Males	**0.70**	0.71	**0.74**	**0.70**	**0.75**	**0.79**

Table 9: Correlation results with real world gender demographic groups based on diverse names for Gemma-2-9b across grades. All values statistically significant with p<0.01.

Model	(r, PC) ↑			(ρ, SC) ↑			AUC ↑		
	G4	G8	G12	G4	G8	G12	G4	G8	G12
Gemma-2-9b	**0.69**	0.61	0.74	**0.69**	0.59	0.73	0.84	0.78	0.84
Gemma-2-27b	**0.69**	**0.68**	**0.78**	**0.68**	**0.67**	**0.78**	**0.85**	**0.83**	**0.9**
Gemma-3-4b	0.58	0.57	0.68	0.57	0.54	0.66	0.77	0.77	0.83
Gemma-3-12b	0.66	**0.70**	0.76	**0.67**	**0.70**	0.75	0.82	0.82	**0.88**
Gemma-3-27b	**0.70**	**0.72**	0.75	**0.74**	**0.72**	0.75	**0.88**	**0.85**	**0.89**
Llama-3-8b	0.42	0.44	0.44	0.40	0.46	0.48	0.67	0.72	0.75
Llama-3-3-70b	0.56	0.53	0.46	0.54	0.52	0.46	0.73	0.74	0.75
Qwen2.5-14B	0.59	0.61	0.58	0.61	0.61	0.57	0.80	0.79	0.80
Qwen2.5-32B	0.60	0.62	0.51	**0.64**	**0.65**	0.50	0.82	0.81	0.76
Qwen3-4b	0.60	0.58	0.62	**0.64**	0.63	0.62	0.79	0.80	0.83

Table 10: Full Correlation and AUC Results of Simulated Students by Model and Grade. All correlations are statistically significant at $p < 0.01$. r = Pearson correlation; ρ = Spearman correlation; AUC = Area Under ROC Curve

Names Used	Grade 4		Grade 8		Grade 12	
	(r, PC) ↑	(ρ, SC) ↑	(r, PC) ↑	(ρ, SC) ↑	(r, PC) ↑	(ρ, SC) ↑
Diverse Asian Names	$0.502_{\pm 0.051}$	$0.474_{\pm 0.030}$	$0.513_{\pm 0.018}$	$0.509_{\pm 0.024}$	$0.572_{\pm 0.037}$	$0.562_{\pm 0.029}$
Diverse Black Names	$0.536_{\pm 0.052}$	$0.495_{\pm 0.023}$	$0.513_{\pm 0.026}$	$0.514_{\pm 0.021}$	$0.585_{\pm 0.028}$	$0.559_{\pm 0.019}$
Diverse Hispanic Names	$0.503_{\pm 0.049}$	$0.491_{\pm 0.049}$	$0.519_{\pm 0.030}$	$0.516_{\pm 0.031}$	$0.580_{\pm 0.027}$	$0.567_{\pm 0.029}$
Diverse White Names	$0.575_{\pm 0.054}$	$0.499_{\pm 0.022}$	$0.514_{\pm 0.019}$	$0.487_{\pm 0.020}$	$0.622_{\pm 0.026}$	$0.578_{\pm 0.030}$
Diverse Names	$\mathbf{0.596}_{\pm 0.027}$	$\mathbf{0.557}_{\pm 0.038}$	$\mathbf{0.582}_{\pm 0.028}$	$\mathbf{0.554}_{\pm 0.034}$	$\mathbf{0.639}_{\pm 0.049}$	$\mathbf{0.612}_{\pm 0.047}$

Table 11: Correlations across different name conditions and grades with standard deviations. r = Pearson correlation, ρ = Spearman correlation. Values shown as $\text{mean}_{\pm \text{SD}}$. We simulate smaller classroom sizes to ensure unique name assignment, with each student sampled from different skill levels. Within each race condition (Asian, Black, Hispanic, White), names are sampled to ensure diversity in gender (male/female) while maintaining racial homogeneity. For example, the Asian condition uses different Asian male and female names across students. The "Diverse Names" condition samples names across all four racial groups while maintaining gender diversity and unique skill-level representation. The lower correlations compared to Table 5 reflect the reduced classroom size, yet the pattern persists: diverse names across racial groups consistently outperform racially homogeneous conditions, showing that cross-racial diversity improves correlations

Model	(r, PC) ↑			(ρ, SC) ↑		
	G4	**G8**	**G12**	**G4**	**G8**	**G12**
Gemma-2-9b	0.003	0.017	-0.120	0.058	-0.048	-0.067
Gemma-2-27b	-0.042	-0.094	-0.018	-0.003	-0.091	-0.031
Gemma-3-4b	0.059	-0.079	0.045	0.053	-0.097	-0.035
Gemma-3-12b	0.020	0.014	0.108	0.006	0.010	-0.105
Gemma-3-27b	0.021	-0.074	-0.059	0.021	-0.087	-0.053
Llama-3-8b	-0.011	0.031	-0.066	-0.037	0.027	-0.048
Llama-3.3-70b	0.095	0.029	-0.014	0.098	0.036	-0.053
Qwen2.5-14B	0.071	0.019	-0.034	0.062	0.022	-0.029
Qwen2.5-32B	0.100	-0.042	-0.050	0.137	0.002	-0.068
Qwen3-4b	0.010	-0.058	0.041	0.022	-0.047	0.024
Qwen3-32b	-0.021	0.018	-0.017	0.028	-0.003	0.020

Table 12: Direct Prompting Correlation Results by Model and Grade. r = Pearson correlation; ρ = Spearman correlation; PC = Pearson Correlation; SC = Spearman Correlation

Model	(r, PC) ↑			(ρ, SC) ↑		
	G4	**G8**	**G12**	**G4**	**G8**	**G12**
Gemma-2-9b	0.028	-0.002	-0.139	0.041	-0.017	-0.108
Gemma-2-27b	-0.018	-0.080	-0.015	-0.016	-0.063	-0.035
Gemma-3-4b	0.031	-0.001	-0.018	0.000	-0.035	-0.039
Gemma-3-12b	0.033	0.027	-0.110	0.022	0.016	-0.124
Gemma-3-27b	0.001	-0.079	-0.037	-0.011	-0.084	-0.016
Llama-3-8b	-0.015	0.037	-0.029	-0.051	0.016	-0.048
Llama-3.3-70b	0.087	0.021	-0.038	0.087	0.046	-0.039
Qwen2.5-14B	0.077	-0.021	-0.048	0.066	-0.027	-0.055
Qwen2.5-32B	0.094	-0.059	0.006	0.117	-0.040	0.000
Qwen3-4b	-0.008	-0.018	0.054	0.003	-0.033	0.044
Qwen3-32b	-0.018	-0.029	0.021	-0.006	-0.044	0.020

Table 13: Averaged Prompting Correlation Results by Model and Grade. r = Pearson correlation; ρ = Spearman correlation; PC = Pearson Correlation; SC = Spearman Correlation

Grade	Classification	Spearman Correlation (ρ) ↑									
		G2-9	**G2-27**	**G3-4**	**G3-12**	**G3-27**	**L3-8**	**L3-70**	**Q2.5-14**	**Q2.5-32**	**Q3-4**
4	Algebra	0.74	0.78	0.81	0.80	0.78	0.54	0.67	0.66	0.74	0.67
	Data Analysis	0.62	0.50	0.51	0.54	0.77	0.10	0.36	0.52	0.37	0.36
	Geometry	0.74	0.83	0.73	0.71	0.59	0.40	0.59	0.67	0.40	0.58
	Measurement	0.81	0.75	0.69	0.78	0.80	0.60	0.72	0.72	0.75	0.63
	Number Properties	0.59	0.61	0.36	0.48	0.60	0.35	0.47	0.45	0.55	0.62
8	Algebra	0.45	0.60	0.35	0.61	0.64	0.30	0.49	0.47	0.56	0.43
	Data Analysis	0.69	0.68	0.59	0.78	0.84	0.59	0.56	0.58	0.60	0.56
	Geometry	0.45	0.57	0.45	0.63	0.64	0.34	0.50	0.54	0.49	0.56
	Measurement	0.77	0.79	0.71	0.78	0.77	0.47	0.61	0.80	0.73	0.69
	Number Properties	0.65	0.71	0.60	0.72	0.76	0.47	0.54	0.60	0.66	0.63
12	Algebra	0.64	0.71	0.69	0.70	0.70	0.53	0.37	0.47	0.35	0.58
	Data Analysis	0.80	0.80	0.72	0.89	0.74	0.38	0.29	0.74	0.55	0.62
	Geometry	0.57	0.61	0.42	0.51	0.69	0.05	0.43	0.21	0.55	0.13
	Measurement	0.87	0.86	0.59	0.81	0.88	0.49	0.53	0.77	0.71	0.65
	Number Properties	0.80	0.86	0.72	0.76	0.73	0.52	0.56	0.58	0.58	0.77

Table 14: Spearman Correlations by Content Area and Grade Level Across Models. Model abbreviations: G2/G3 = Gemma-2/3, L3 = Llama-3, Q2.5/Q3 = Qwen2.5/3. Numbers indicate parameter count in billions (e.g., G2-9 = Gemma-2-9b). To understand where simulations succeed and fail, we analyze correlations across the five content areas. From the table, we observe that problems that evaluate students understanding on the Measurement content area, consistently achieve strongest correlations. Problems on Algebra and Geometry show the weakest performance: Algebra averages ρ=0.58 across grades with particularly poor Grade 8 performance (most models), while Geometry shows high variance with weaker correlations at Grade 12 for some models (Llama-3-8B).

Grade	Classification	Distractor Match (%) ↑									
		G2-9	**G2-27**	**G3-4**	**G3-12**	**G3-27**	**L3-8**	**L3-70**	**Q2.5-14**	**Q2.5-32**	**Q3-4**
4	Algebra	43.3	43.3	26.7	33.3	43.3	46.7	36.7	50.0	36.7	40.0
	Data Analysis	44.0	48.0	40.0	56.0	48.0	44.0	40.0	48.0	48.0	28.0
	Geometry	36.8	36.8	31.6	31.6	42.1	31.6	31.6	36.8	21.1	52.6
	Measurement	27.8	46.3	46.3	38.9	29.6	51.9	50.0	42.6	40.7	53.7
	Number Properties	34.0	29.0	34.0	32.0	20.0	40.0	43.0	30.0	37.0	38.0
8	Algebra	27.8	29.2	26.4	29.2	45.8	22.2	30.6	27.8	33.3	33.3
	Data Analysis	33.3	42.4	30.3	30.3	30.3	27.3	18.2	45.5	30.3	42.4
	Geometry	28.9	31.1	40.0	35.6	28.9	31.1	35.6	33.3	26.7	35.6
	Measurement	41.7	52.1	33.3	45.8	35.4	37.5	37.5	25.0	45.8	35.4
	Number Properties	34.5	38.1	36.9	42.9	45.2	27.4	29.8	42.9	44.0	28.6
12	Algebra	31.4	25.7	40.0	42.9	48.6	25.7	17.1	25.7	25.7	34.3
	Data Analysis	42.1	42.1	36.8	42.1	26.3	10.5	10.5	42.1	31.6	42.1
	Geometry	50.0	50.0	42.9	35.7	21.4	35.7	50.0	14.3	35.7	35.7
	Measurement	26.1	30.4	43.5	39.1	39.1	47.8	34.8	43.5	34.8	21.7
	Number Properties	36.7	46.7	40.0	50.0	46.7	20.0	23.3	50.0	50.0	26.7

Table 15: Distractor Match by Content Area and Grade Level Across Models. Model abbreviations: G2/G3 = Gemma-2/3, L3 = Llama-3, Q2.5/Q3 = Qwen-2.5/3. Numbers indicate parameter count in billions.

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 16: 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 17: Sample student responses from Gemma-2-9B across skill levels for Grade 8 question item.

Question	**Which of the following expressions does NOT have the same value as two times five times eight?**
Answer Choices	A. two times forty; B. ten times sixteen; C. five times eight times two; D. two times five times two times four
Content Area	Number Properties and Operations
Below Basic Response	{'**reasoning1**': 'I think that is fifty minus twelve...', 'answer key': 'B', '**reasoning2**': 'I know two times five times eight equals forty. I think two times forty is the only choice that doesn't have forty in it.', 'answer key': 'B'}
Basic Response	{'**reasoning1**': 'I know 2 times 5 is 10, and 5 times 8 is 40, so I thought 10 times 40. I see the answer is 200, so I thought I did something wrong, and I checked the other answers. I didnt see any that were really wrong, but I decided to pick the one that was different. It was the only one that had the word times twice, so I picked that one', 'answer key': 'D', '**reasoning2**': 'I know two times five times eight equals forty. I think two times forty is the only choice that doesn't have forty in it.', 'answer key': 'B'}
Proficient Response	{'**reasoning1**': 'I know that 2 times 5 times 8 is 80, so I need to figure out which one is not 80. I know that 2 times 40 is 80, so that's not the answer. I know that 10 times 16 is 160, so that's not the answer. I know that 5 times 8 times 2 is 80, so that's not the answer. I know that 2 times 5 times 2 times 4 is 80, so that's not the answer. ', 'answer key': 'B', '**reasoning2**': 'I know that 2 times 5 times 8 is 80, so I need to figure out which one is not 80. I know that 2 times 40 is 80, so that's not the answer. I know that 10 times 16 is 160, so that's not the answer. I know that 5 times 8 times 2 is 80, so that's not the answer. I know that 2 times 5 times 2 times 4 is 80, so that's not the answer. I think the answer is A.', 'answer key': 'B'}
Advanced Response	{'**reasoning1**': 'The first step is to find two times five, which is ten. Then, I can multiply ten by eight to get eighty. Since two times five times eight is equal to eighty, I will eliminate any answer choices that equal eighty. The only answer choice that does not equal eighty is two times forty, so I will choose this answer. I also notice that the other answer choices can be simplified to have factors of five and eight, so this choice must be the right one. I could also double-check by plugging in numbers, but this would take longer.', 'answer key': 'A', '**reasoning2**': 'I can see that two times five times eight is 40, so I need to find the expression that doesn't also equal 40. I notice that ten times sixteen is 160. So it is not the same as the other three choices, which are 40, 80, and 80. So my answer is B.', 'answer key': 'B'}

Table 18: Sample student responses from Llama-3-8B across skill levels for Grade 4 question item. This model selects B despite flawed reasoning. Below Basic claims "2×5×8=40" (wrong =80); Proficient calculates correctly, states "answer is A," yet picks B. This model prioritizes accuracy over role-fidelity, generating superficial errors while making correct selections.