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- and 8th-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 the simulated student scores align to an extent closely with real student success. We also observe that the individual models contribute different strengths and combining them could improve the correlation compared to using the individual models in some cases.

1 Introduction

Math word problems (MWPs) are a common instrument of student evaluation as well as 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). For students of differing cultural backgrounds, math word problems that require access to culturally specific knowledge may threaten the validity of these items as an assessment tool, and introduce barriers to learning for students who may already face other disadvantages. Thus the 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 either in generating these question items (Singh et al., 2021) or relying on their retrospective student performance data after analyzing student performance using psychometric methods of evaluations(Harris, 1989; Bond and Fox, 2013).

Recent works, shows 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: Generative-Student profiles built from knowledge components detect hard items without real data (Lu and Wang, 2024), the Classroom Simulacra framework which models full classroom dynamics (Xu et al., 2025), 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. Our approach grounds these simulations against real student performance data using standardized psychometric techniques (e.g., Rasch modeling), validating their predictive power in estimating item difficulty accurately . 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 mirror those provided by test developers?

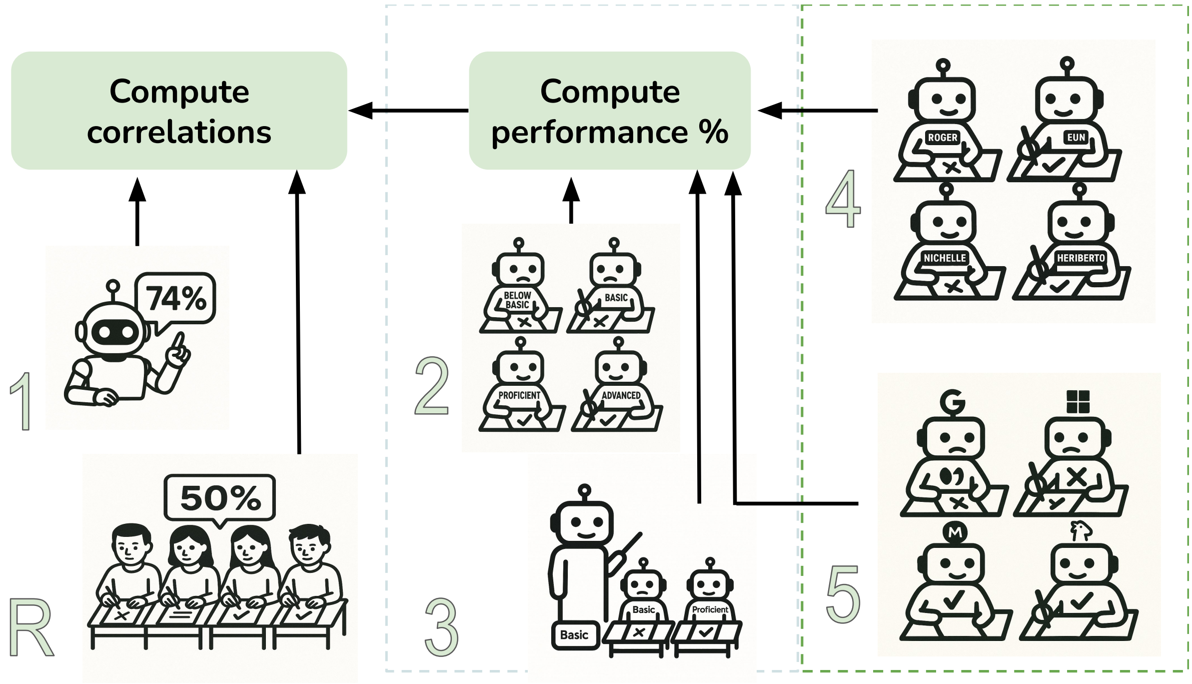

Figure 1: **Our simulation pipeline.** We estimate real-world math word problem difficulty, R using LLM-simulated students. We implement: (1) Direct percentage estimation of correct answers; (2) Student role-play across skill levels; (3) Teacher-based predictions for students at different skill levels; (4) First name + student simulation; and (5) Skill-mapped ensemble with different LLMs representing distinct skill levels. Correlation with actual student performance evaluates each method's accuracy in predicting item difficulty.

2 Preliminaries

2.1 Data

We collect 79 Multiple Choice Math Word Problems(MWP) from the National Report Card website (National Center for Education Statistics (NCES)) for grades 4 and 8. 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. We use these as it provides actual student performance statistics across the nation, providing a good 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. These questions from the NAEP has also gone through a rigorous development and validation processes. Each problem includes a question, answer choices, correct answer and meta-data (difficulty, content area, grade level, and student performance statistics across demographic groups including gender). Table 6 summarizes the distribution of problems by grade level and difficulty. Despite the relatively small size of the data, the selected MWPs cover a diverse range of mathematical concepts. Table 7 presents the distribution of problems across content areas. We limited our scope to grades 4 and 8 word problems, filtering out visual or diagram-based problems that would require different LLM capabilities beyond our current research scope. These math word problems are used for evaluation, not training, thus reducing concerns about model overfitting to the test set. Given that we are not conducting combinatorial experiments across multiple variables but rather performing a straightforward evaluation of LLM capabilities against human benchmarks, this sample size provides sufficient coverage of key mathematical concepts while remaining manageable.

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 word problem p with multiple choices $a_1, a_2..a_m$. We compute the performance as an estimated proportion $\hat{y}_p \in [0, 1]$ representing the predicted percentage of students who would correctly answer problem p. Our goal is that given a set of math problems $P = p_1, p_2, ..., p_m$ 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 relative difficulty of the word problems. 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.

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 student to 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 generate simulated students based on diverse demographics and grade levels by using a prompt template that includes "[NAME]" and "[GRADE]" placeholders. By substituting these placeholders with first names statistically associated with specific racial/ethnic and gender identities and the grade level, we derive demographic information directly from the assigned name. 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 enrich the simulation process, we hypothesize simulating more diverse students could lead to better population-level difficulty estimates. To this end, we extend the idea from different NLP studies that have used first names as proxies for different demographics attributes (Caliskan et al., 2017; Acquaye et al., 2024; Sancheti et al., 2024; Zhang et al., 2024). We use first names as a proxy for this demographic information to simulate diverse students. We select 48 names that are most representative of four races/ethnicities (Asian, Black, Hispanic and White), distributed evenly across two genders (female and male). These names were selected based on their usage in (Sancheti et al., 2024; An et al., 2024). Each intersectional demographic group has six names, totaling 48 names. A comprehensive list of these names can be found in appendix A.3.

Models We experiment with open-source LLMs of varying sizes, including Llama-3.1-70B (Dubey et al., 2024), Phi-3.5-mini (Abdin et al., 2024) and Mixtral-8x7B (Jiang et al., 2024), based on math benchmark performance. We evaluate these models in zero shot prompting strategy to answer the 79 questions to get the models accuracy as the models knowledge can constraint its ability to correctly simulate certain skill levels for the simulation. Aside from Phi in Table 2 (whose Grade-8 accuracy is 0.61), all other models achieved good baseline performance (> 0.77–1.00), indicating they have enough subject knowledge to answer the math word problems correctly before being adapted to student-level simulation.

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 math word problem correctly with prompt A.2. This way, we get the measure of the model's understanding of the question's difficulty from a predicted percentage of students who would answer correctly. 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 also include the description of the class size with the number of students in the different skill levels, in addition to the grade level in another experiment. 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 uses greedy decoding to generate a single deterministic output by selecting the highest-probability token at each step. The second approach uses stochastic sampling (temperature $T = 0.3$) to generate three responses, then aggregates their predictions by averaging the resulting probabilities. We anticipate that simulating multiple students 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

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

Table 1: Direct Prompting Correlation Values by Grade and Model

Grade		Gemma-2-9b-it		Phi-3.5-mini		Mixtral-8x7B		Llama-3.1-70B	
		Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
4	Only Grade	0.204	0.250	-0.072	-0.172	0.059	0.079	0.025	0.032
	Only Grade (Averaged)	0.227	0.194	-0.139	-0.030	0.161	0.244	0.230	0.209
	Grade + Class Information	0.170	0.294	0.118	0.067	-0.067	-0.073	0.163	0.191
	Grade + Class Information (Averaged)	0.361	0.308	0.033	-0.040	-0.106	0.034	0.085	0.076
8	Only Grade	0.548	0.488	0.111	0.175	0.352	0.342	0.223	0.158
	Only Grade (Averaged)	0.483	0.348	0.308	0.344	0.410	0.313	0.408	0.479
	Grade + Class Information	0.137	0.117	0.308	0.266	0.284	0.342	0.269	0.258
	Grade + Class Information (Averaged)	0.061	0.087	0.279	0.277	0.328	0.250	0.026	0.327

Table 2: LLM accuracy on NAEP math word problems by grade level

Model	**Grade 4**	**Grade 8**	**Total**
Gemma-2-9b-it	0.86	0.83	0.85
Phi-3.5-mini	0.72	0.61	0.67
Mixtral-8x7B	0.77	0.69	0.73
Llama-3.1-70B	1.00	0.81	0.91

difficulty estimates.

Simulated Classroom Performance Estimation We prompt the LLMs in three role-play prompts variants, each of which generates N simulated student responses per question. The student prompt asks the LLM to answer the question as a student of a given skill profile with prompt A.3, while the teacher prompt asks the LLM to role play a teacher who can predict a given student of a skill profile response with prompt A.4. We also prompt the model as a student with a first name and a skill profile with A.5. For a given question, we first compute, 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 79 items to get the success rate of the classroom.

Model Ensembling Classroom Performance Estimation We explore model diversity as a dimension to simulate diverse student classroom, by aggregating the outputs of all LLMs. With this, we can get more accurate and robust estimates than relying on one single model's prediction (Mehri and Eskénazi, 2019; Page et al., 2023; Mangalvedhekar et al., 2023). We ensemble these LLMs outputs in an averaging and a skill mapping strategy. In the averaged ensemble approach, for each of the four LLMs, we sample N student responses and calculate each model's simulated percentage correct and then average these values across the models to get a final averaged percentage correct. In the skill-mapped ensemble approach, we assign exactly one model to each skill bucket and generate responses for students of that particular skill level. For example, we can have LLM 1 simulating students who are Below Basic, LLM 2 simulating students who are Basic, LLM 3 simulating students who are Proficient, and LLM 3 simulating students who are Advanced. We aggregate these responses and estimate the percentage correct value for the class of N size.

Rasch IRT Difficulty Estimation We estimate item difficulties using a Rasch IRT model (Rasch, 1980) fitted to binary response data simulated from our best model, with relatively higher correlations with the real world students, gemma-2-9b.

$$P(X_{ni} = 1 \mid \theta_n, b_i) = \frac{\exp(\theta_n - b_i)}{1 + \exp(\theta_n - b_i)}$$

where:
θ_n = Ability of student n; b_i = Difficulty of item i
This model provides the difficulty estimates for each item on a latent scale. We simulated student responses from large language models for Grade 4 and Grade 8 math items across four skill buckets (Below Basic, Basic, Proficient, Advanced). Each simulated student generated responses was graded on a binary (1=correct, 0=incorrect) answers. We fit these responses to a Rasch model, which simultaneously estimated each item's difficulty and each student's ability. To test whether these difficulties align with difficulty categories, we employed k-means clustering (with k=3) on the Rasch-estimated difficulties. The resulting clusters were grouped based on these numeric difficulties into categories—Low, Medium, High.

3.2 Evaluation

For each math problem, we compute an accuracy, representing the percentage of students in a simulation that got the question right, which we compare

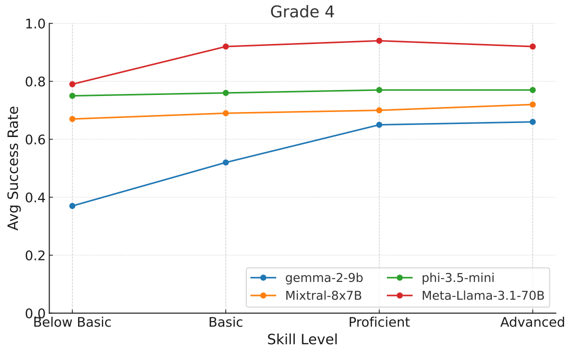

Figure 2: Average Simulated Accuracy by Skill Level Across LLMs for Grade 4

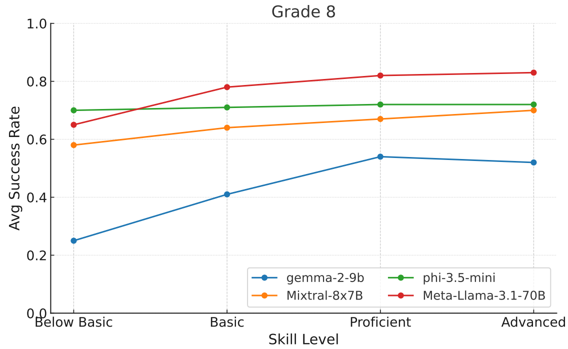

Figure 3: Average Simulated Accuracy by Skill Level Across LLMs for Grade 8

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 problems, we use Pearson (Pearson) and 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 problems 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. We vary the class size to establish an optimal efficiency-performance trade-off. While increased simulation sizes potentially reduces sampling variance, they also incur higher computational costs. By testing multiple classroom sizes (40, 100, and 300 students), we aim to identify the point at which additional simulated responses no longer yield substantial improvements in correlation with actual performance data to have a balanced trade off.

4 Results and Discussion

4.1 Simulated Students' Accuracy are Aligned with the Student Skill Profiles

As expected in the real world, the simulated students accuracy progressively improves as the skill levels increase from Below Basic to Advanced as shown in Figures 2 and 3, yet the sharpness of that gradient varies by LLM. Llama-3 70B shows the highest performance, in Grade 8 its Advanced group outscores its Below Basic group on average. Gemma-2 9B also tracks the four skill buckets reliably but with a slightly compressed gap, while Mixtral-8×7B collapses Basic and Proficient into almost identical curves, and Phi-3.5-mini often overestimates low-skill performance, which produces an almost constant gradient. In short, the simulated students' answers do align with their scripted skill levels, but the fidelity of that varies by LLM.

4.2 Simulated Students Performance Correlates with Real-World Student Performance with Varying Fidelity

Direct Estimation For Grade 4, in the baseline results in Table 1, the correlations are generally moderate(often around 0.2 or lower), which suggests that LLMs struggle to capture the difficulty faced by younger students. For Grade 8, although some models like gemma-2-9b-it show relatively higher correlations, the overall inconsistency across models and conditions remains evident. Particularly, the 'Only Grade' baseline outperforms the 'Grade + Class Information' baseline, suggesting that adding more information in this prompting approach, does not necessarily improve the predicted percentage. Also, comparing a single greedy decoding response with an averaged result from multiple responses shows variability; while averaging may smooth out individual anomalies, it obscures the fundamental instability of the model's predictions indicating that simply relying on direct prompting is insufficient for accurately predicting student performance.

Simulated Estimation Using our defined sampling proportions, the simulated student responses show some correlation with real-world performance as seen in Table 3, although the strength varies across models and grades. Specifically, Gemma-2 9B

Table 3: Correlation between 100 simulated students and real world student performance

Model	Grade	Student		First Name + Student	
		Pearson	Spearman	Pearson	Spearman
Gemma-2-9b-it	4	0.76	0.79	0.74	0.77
	8	0.74	0.77	0.72	0.73
Phi-3.5-mini	4	0.45	0.50	0.57	0.61
	8	0.53	0.55	0.61	0.64
Mixtral-8x7B	4	0.39	0.42	0.54	0.54
	8	0.63	0.64	0.52	0.57
Llama-3.1-70B	4	0.57	0.60	0.71	0.72
	8	0.54	0.58	0.57	0.60

consistently achieves higher correlations for both grades , reflecting its ability to simulate performance differences effectively across skill levels, as observed in the clear skill gradients reported earlier. In contrast, Phi-3.5-mini and mixtral exhibits weaker correlations likely due to its struggles in accurately distinguishing between skill levels, particularly overstating lower skill performances. Notably, adding first names slightly improves correlations, in the Phi-3.5-mini model and Llama model, indicating how demographic contextualization could be a tool to boost simulation. Thus, explicitly modeling student diversity via names can boost the correlation with actual student outcomes.

We also see in Table 4 the correlations between simulated and real-world student performance under the student and teacher approach, using different simulated class sizes. Notably, sampling classes with larger class sizes, achieving a correlation of 0.791 at Grade 4 and 0.77 at Grade 8 for the largest class size of 100. This further increased for a class size of 1000 to a correlation of 0.82 for grade 4 and 0.79 for grade 8. Considering the correlations do not improve significantly, we continue our simulations with 100 students. Also, prompting the model as a student got better correlations compared to asking the LLM to role a student. This difference suggests that role playing a student may better align with the Gemma, as student prompts more naturally simulate varied response patterns.

Ensembling Estimation We use the mapping ensemble: Gemma-2-9b-it answers as Below Basic students, Mixtral-8x7B as Basic, Phi-3.5-mini as Proficient, and Llama-3.1-70B as Advanced. We derive this mapping by noting that each model's accuracy peaks at a different skill levels in the plots in Figures 2 and 3—Gemma lowest, Mixtral next, Phi mid, Llama highest—so we assign them to those matching skill groups. The mapping ensemble achieves a slightly higher Grade 4 correlation (around 0.80 Spearman), slightly outperforming the averaged ensemble approach. Grade 8 sees a similar pattern: the skill-mapped ensemble still outpaces any individual model, with a modest correlation gain over the averaged ensemble.

Table 4: Correlation Values for Gemma-2-9b-it by Grade, Class Size, and Prompt Approach

Grade		Student		Teacher	
		Pearson	Spearman	Pearson	Spearman
4	40	0.75	0.78	0.684	0.740
	100	0.76	0.79	0.70	0.75
	300	0.78	0.81	0.70	0.75
8	40	0.73	0.76	0.65	0.65
	100	0.74	0.77	0.65	0.65
	300	0.76	0.78	0.65	0.66

Table 5: Correlation between ensembled models simulated students and real world student performance

Grade	Averaged		Mapping	
	Pearson	Spearman	Pearson	Spearman
4	0.72	0.75	0.78	0.80
8	0.62	0.58	0.71	0.72

4.3 Simulated Student Performance is a good indicator of the difficulty of a question

By overlaying expert-assigned difficulty labels from the meta data from NAEP, we visually confirmed that most items aligned closely with expectations for Grades 4 in Figure 4: Easy items predominantly clustered in the Low difficulty group, Medium items in the Medium group, and Hard items in the High group. While a few items were misaligned, this overall consistency provides evidence that our LLM simulations simulates real student responses and, thus, can serve as a tool for approximating item difficulty. We however observe for Grade 8 in Figure 4, a slightly less perfect alignment which signals that additional calibration is needed to improve item-level fidelity at the Grade 8 level.

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.Lu and Wang (2024) introduce a *Generative Students* framework where GPT-4 is

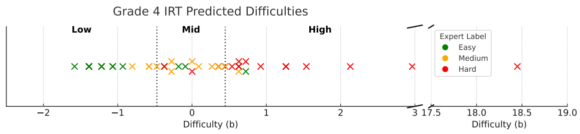

Figure 4: Grade 4 IRT predicted difficulties clustered and visualized with actual difficulty from real world data

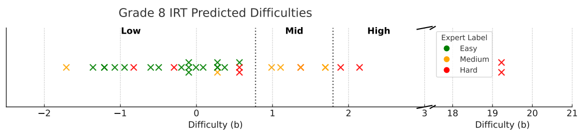

Figure 5: Grade 8 IRT predicted difficulties clustered and visualized with actual difficulty from real world data

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 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 and demographic cues. The correlation further improves when model diversity is exploited—ensembling LLMs based on their relative strengths across skill levels produces richer, more consistent performance estimations. This informs directions for future work, exploring multiple model variants and ensemble methods to capture more diverse students through multiple prompting dimensions—skill, name, socioeconomic background to ensure more stable predictions. With this, we can run formal fairness analyses (such as differential item functioning) to systematically verify that difficulty flags affect all student groups equitably.

Limitations

LLM Limitations Our experiments relied on four 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 can also provide more robust conclusions. Additional evaluations would enhance generalization.

Limited Data size We evaluated model-generated responses on 79 multiple-choice questions for Grade 4 and Grade 8. While these cover a range of difficulty levels and content areas, the size and scope of the questions remain constrained.

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 and Grade 8 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.

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.

Acknowledgements

References

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

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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.

Sudeep Mangalvedhekar, Kshitij Deshpande, Yash Patwardhan, Vedant Deshpande, and Ravindra Murumkar. 2023. Mavericks at araieval shared task: Towards a safer digital space - transformer ensemble models tackling deception and persuasion. In *ARABICNLP*.

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.

Shikib Mehri and Maxine Eskénazi. 2019. Multi-granularity representations of dialog. *ArXiv*, abs/1908.09890.

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

Saurabh Page, Sudeep Mangalvedhekar, Kshitij Deshpande, Tanmay Chavan, and Sheetal S. Sonawane. 2023. Mavericks at blp-2023 task 1: Ensemble-based approach using language models for violence inciting text detection. *ArXiv*, abs/2311.18778.

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.

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.

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.

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*.

Difficulty Level	Total #	Grade 4 #	Grade 8 #
Easy	33	15	18
Medium	21	13	8
Hard	25	15	10
Total	**79**	**43**	**36**

Table 6: Breakdown of question difficulty by grade.

Content Area	Count
Number properties and operations	39
Measurement	16
Algebra	10
Data analysis, Statistics, and Probability	8
Geometry	6

Table 7: Distribution of content areas being tested in the dataset.

Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. 2025. 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*.

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?

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: Teacher Simulation Prompt

Task:
You are an expert, experienced math instructor that can reliably predict how a Below Basic student in the {grade}th grade will 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 but rather this expert, experienced math instructor that can predict how a {skill level} student will answer the math problem, and completely simulate yourself as one.

Prompt A.5: 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

- Phi: https://ai.meta.com/llama/license/
- Llama3: https://llama.meta.com/llama3/license/
- Mistral: https://mistral.ai/terms-of-service/
- 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.

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