Can LLMs Reliably Simulate Real Students' Abilities in Mathematics and Reading Comprehension?

KV Aditya Srivatsa **Kaushal Kumar Maurya** **Ekaterina Kochmar**
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{vaibhav.kuchibhotla, kaushal.maurya, ekaterina.kochmar}@mbzuai.ac.ae

Abstract

Large Language Models (LLMs) are increasingly used as *proxy students* in the development of Intelligent Tutoring Systems (ITSs) and in piloting test questions. However, *to what extent these proxy students accurately emulate the behavior and characteristics of real students remains an open question.* To investigate this, we collected a dataset of 489 items from the National Assessment of Educational Progress (NAEP), covering mathematics and reading comprehension in grades 4, 8, and 12. We then apply an *Item Response Theory (IRT)* model to position 11 diverse and state-of-the-art LLMs on the same ability scale as real student populations. Our findings reveal that, without guidance, strong general-purpose models consistently outperform the average student at every grade, while weaker or domain-mismatched models may align incidentally. Using grade-enforcement prompts changes models' performance, but whether they align with the average grade-level student remains highly model- and prompt-specific: no evaluated model–prompt pair fits the bill across subjects and grades, underscoring the need for new training and evaluation strategies. We conclude by providing guidelines for the selection of viable proxies based on our findings.[1]

1 Introduction

Large language models (LLMs) are capable of generating fluent and coherent text and excelling at many complex tasks (Chang et al., 2024; Zhao et al., 2024). Their rise offers new opportunities for educational technology, notably in (i) intelligent tutoring systems (ITS; Wang et al., 2024) and (ii) *piloting assessments* before they go live (Liu et al., 2025; Grohs et al., 2024). ITS provides targeted feedback and adaptive instruction, while reliable assessments track learning without bias. Yet both require understanding how real students would interact with them, which is extremely challenging to verify.

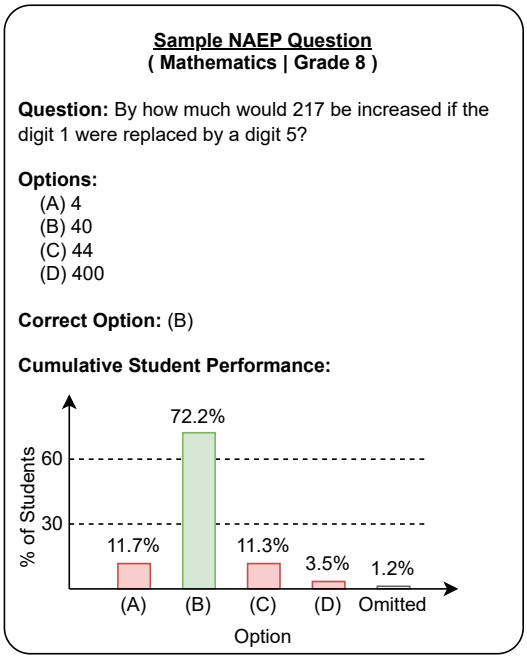


Figure 1: Sample NAEP question from grade 8 mathematics.

Ideally, tutors and test forms should be vetted on representative student samples across skill levels. This, however, is resource-intensive, especially in regions already short on teachers and infrastructure (UNESCO, 2023; Woolf et al., 2013). Teacher-led evaluations (e.g., Macina et al., 2023) and static logs similarly fail to scale or capture the dynamics of new items and adaptive strategies (Belz et al., 2023; U.S. Department of Education, 2023). These constraints motivate alternatives that enable rigorous, equitable evaluation at scale.

An emerging approach is for *simulate* students with LLMs (Mollick et al., 2024; Sonkar et al., 2024). Proxy models can be conditioned on at-

[1]All related code and data is available at https://github.com/kvadityasrivatsa/IRT-for-LLMs-as-Students

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 988–1001

tributes such as grade level, offering fast, repeatable tests of tutor features or item quality. However, current evaluations are based on an expert judgment of plausibility (Macina et al., 2023), leaving open the question of how closely such proxies match real student performance. Similarly, in psychometrics, LLMs have been used as *synthetic examinees*: e.g., Liu et al. (2025) show that `GPT-3.5/4` answer sets yield item statistics that mirror a 50-student pilot, and Grohs et al. (2024) demonstrate that ChatGPT can pre-flag weak or biased items. However, these studies treat LLMs as single test-takers and do not look into whether persona prompts can tie their abilities to specific grade bands.

Our approach: We apply IRT (Baker, 2001) to measure how 11 diverse LLMs and real students perform on the same grade-level questions. Using data from the National Assessment of Educational Progress (NAEP) (National Center for Education Statistics, 2022) for mathematics and reading in grades 4, 8, 12, we evaluate whether the LLM responses (both *generic* and *grade-conditioned*) align with authentic student response patterns. Specifically, we address the following research questions: **RQ1** – Under standard prompting, how do LLMs compare with real students across grades and subjects? **RQ2** – When asked to act as an average student in a given grade: How does LLM performance change? (**RQ2.1**) Does the shift match real grade-level patterns? (**RQ2.2**)

The main contributions of our work are as follows:

- We compile and release a dataset[2] sourced from NAEP of real student responses to subject-specific, grade-targeted questions, covering two subjects (mathematics and reading assessment) and three grade levels (4, 8 and 12).
- We adapt Item Response Theory (IRT) to assess the alignment between LLM-generated responses and actual student performance patterns.
- We conduct an evaluation of 11 diverse LLMs, examining how well they approximate student responses under both *unenforced* (generic) and *grade-enforced* prompts.

[2]https://github.com/kvadityasrivatsa/IRT-for-LLMs-as-Students

2 Related Work

Simulated Students in Intelligent Tutoring Systems Early simulated-student work relied on production rule *apprentices* that learn from worked examples and then reproduce step-level behavior inside an ITS. The *SimStudent* / *Apprentice-Learner* family shows that such models can generate realistic error types and serve as policy learners to hint (Matsuda et al., 2023; Smith et al., 2024). More recent studies graft LLMs onto this pipeline: it has been shown that `GPT-4` "think-aloud" traces improve bug-library discovery and fine-grained skill tagging, while LLM agents at dialogue level can populate entire synthetic classroom cohorts (Mollick et al., 2024). These approaches demonstrate that generative text can complement symbolic learner models, yet they rarely test whether the *ability* distribution of synthetic learners matches that of real students – a gap we address through our analysis.

LLM-Generated Responses for Item Calibration Psychometric studies have begun to treat LLM outputs as *synthetic examinee responses*. Liu et al. (2025) show that the `GPT-3.5/4` answer sets yield 3PL item statistics that match a 50-student baseline, reducing the pretest costs. Grohs et al. (2024) use ChatGPT to filter out low information or biased items. He-Yueya et al. (2024) further adapt IRT to align LLM and human response patterns, while Zelikman et al. (2023) simulate K-12 students. However, these works produce only aggregate correlations; they do not examine whether an LLM's *latent ability* aligns with a particular grade band or whether persona prompts shift that ability in predictable ways. Our work closes this gap with an IRT model that maps LLM performance to grade-level performance.

Persona-Conditioned Prompting and Alignment Prompting a model with an explicit role (e.g., "*You are a 4^{th}-grade student*") can change both reasoning depth and surface style. Benedetto et al. (2024) find that a one-sentence student-level prompt lets `GPT-4` imitate weak, average, and strong test-takers across subjects, although adherence to the target level is uneven. Broader evaluations such as CharacterEval (Tu et al., 2024) measure persona consistency in dialogues, while Kim et al. (2024) show that role prompts can either help or hurt accuracy depending on task characteristics. None of these efforts connect persona adherence to *quan-*

titative grade-level ability estimates, nor do they compare default and persona-conditioned ability curves within a unified IRT framework.

Together, these strands indicate that (i) LLMs are already employed as simulated students and psychometric stand-ins, and (ii) persona prompts shift model behaviour without a principled link to grade-level metrics. Our study unifies the two directions by applying an IRT model to quantify how default and persona-conditioned LLM outputs align with average student performance at grades 4, 8, and 12.

3 NAEP Data

3.1 Source and Composition

We prepared our dataset using publicly available items and student response data from the National Assessment of Educational Progress (NAEP) (National Center for Education Statistics, 2022),[3] a large-scale assessment program administered by the National Center for Education Statistics (NCES). NAEP periodically assesses student achievement across the United States in key subject areas, including mathematics and reading. These assessments are conducted in grades 4, 8, and 12, offering a cross-sectional perspective on student proficiency throughout K–12 education.

3.2 Coverage and Educational Context

We source questions from both the *mathematics* and *reading comprehension* assessments at the three grade levels, capturing a broad spectrum of student performance and cognitive development throughout different educational stages. We focus on these two subjects for two reasons: (1) numeracy and literacy are considered fundamental skills (e.g., Williams, 2003); and (2) NAEP data cover three grades for these subjects, while many other subjects only cover one or two grades. Math questions span topics such as measurement, algebra, geometry, and probability and statistics, with overall difficulty scaling with grade level. Reading comprehension items are based on passages whose average length increases with grade. The corresponding questions shift from direct factual queries in lower grades to those requiring interpretation and reflection at higher levels.

Each record contains the original question, multiple choice options, the correct annotated answer, and anonymized aggregate response patterns. For each item, the dataset reports the percentage of students who selected each option or omitted the question. Figures 1 and 4 show representative examples from the grade-8 mathematics and grade-12 reading subsets, respectively.

Since NAEP is a continually administered assessment, this dataset can be periodically updated with newly released items. This makes it a dynamic resource that can evolve along with changes in educational standards and student performance distributions, offering long-term utility for evaluating automated student proxies and similar tasks.

3.3 Preprocessing and Filtering Criteria

NAEP assessments encompass a variety of question types, modalities, and response formats. Given that this is a preliminary effort to develop a quantitative and interpretable framework for aligning LLM performance with real student behavior, the inclusion of diverse modalities can introduce confounding factors that obscure analysis. For example, items that involve images, diagrams, or tables introduce the additional variable of visual comprehension, making it difficult to isolate language understanding as the primary factor in model performance. Similarly, free-form responses present evaluation challenges: gold-standard answers are often limited in number and may not capture the full range of acceptable responses. Assessing these reliably often requires expert judgment, which undermines the feasibility of scalable, LLM-based evaluation.

In contrast, multiple choice questions offer clearly defined answer sets, enabling a more straightforward and objective evaluation, which is crucial for both quantitative benchmarking and interpretability via Item Response Theory (IRT). Consequently, we apply two main filtering criteria when constructing our dataset.

- *Text-only content:* We exclude any items that involve diagrams, tables, or multimedia elements, retaining only questions and instructions presented in text.
- *Multiple-choice format:* We include only multiple choice questions (MCQs), which support standardized evaluation and facilitate downstream processing, such as answer extraction and IRT-based analysis.

After filtering, our final dataset consists of 489 multiple-choice items in English: 249 from mathe-

[3] https://nces.ed.gov/nationsreportcard/

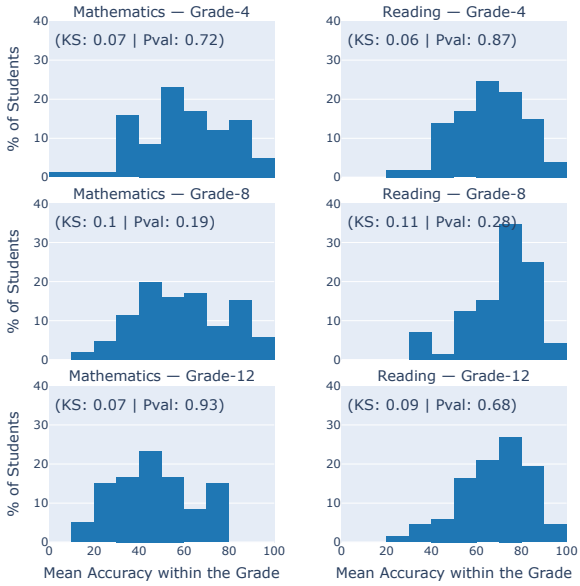


Figure 2: Distribution of question-level accuracy in NAEP assessments across grades and subjects. KS statistics and corresponding p-values are reported to assess normality; distributions with p > 0.05 are considered consistent with a normal distribution.

matics and 240 from reading. Table 3 summarizes key statistics from the dataset. Figure 2 shows the distribution of questions answered by students with varying accuracy for each subject and grade. We calculate the Kolmogorov-Smirnov (KS) statistic for each distribution to test for normality, and all subsets sufficiently (p-value > 0.05) follow a normal distribution.

4 Proposed Methodology

Our goal is to evaluate LLMs alongside human students recorded in the NAEP dataset by estimating their answering ability on a shared scale. To do this, we draw on Item Response Theory (IRT; Baker (2001)), a well-established framework in educational measurement. IRT enables us to jointly model the ability of test takers and the difficulty of individual test items using probabilistic principles.

4.1 Estimating Student (LLM) Ability

We begin with the Rasch model (Rasch, 1960), a widely used and interpretable form of IRT. It assumes that the probability of a correct response depends only on the difference between a participant's ability and an item's difficulty. This model uses a single parameter per item, that is, difficulty b_i, and one ability parameter θ_i per participant.

The Rasch model defines the probability that participant i correctly answers item j as:

$$P(X_{ij} = 1) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}}, \tag{1}$$

where $\theta_i \in \mathbb{R}$ is the ability of the participant i, $b_j \in \mathbb{R}$ is the difficulty of item j, and $\mathbb{R}$ is the common scale of difficulty or ability.

To estimate b_j, we use the empirical proportion p_j of correct responses for each item in the population. A simple approximation is:

$$b_j \approx \log\left(\frac{1 - p_j}{p_j}\right), \tag{2}$$

which reflects that the harder items (with lower p_j) have higher difficulty values (Bond and Fox, 2015). Once the item difficulties are known, the ability of each participant θ_i can be estimated using marginal maximum likelihood or Bayesian inference based on their response pattern.

4.2 Grade-Alignment Prompting

Our first research question (RQ1) investigates how an LLM's problem-solving ability compares to that of average students at different grade levels, specifically grades 4, 8 and 12, based on the NAEP dataset. To measure this, we begin with a minimal zero-shot prompt, which we refer to as **UNENFORCED** (see Appendix Figures 5 and 9 for exact prompt templates). This prompt simply presents the question to the model without any added instructions or persona guidance.

Our second research question (RQ2) explores whether LLMs can align their responses with the answering patterns and performance levels of students in specific grades. To probe this, we design a set of increasingly guided zero-shot prompts that aim to steer the model toward grade-level reasoning.

1. **GRADEENFORCEDMINIMAL**: Identical to the `Unenforced` prompt, but with the added instruction that the model should act as an average student from a specific grade (4, 8, or 12). The exact prompts are presented in Appendix Figures 6 and 10.
2. **GRADEENFORCEDBASICCOT**: Builds on the minimal version by prompting the model to consider what an average student at the specified grade would likely choose and why. This prompt encourages brief, grade-aware reasoning and reflects the student's typical reasoning ability and common error patterns.

See Figures 7 and 11 in the Appendix for the exact prompts.

3. **GRADEENFORCEDFULLCOT**: Adds further scaffolding by dividing the reasoning process into two steps. First, the model is instructed to reflect on whether an average student at the given grade level would be likely to answer the question correctly. Second, based on that reflection, the model either justifies a correct answer or, if the student is unlikely to succeed, selects and explains the most plausible incorrect answer. See Figures 8 and 12 in the Appendix for the exact prompts.

The design of the GRADEENFORCEDBASICCOT and GRADEENFORCEDFULLCOT prompts is inspired by Benedetto et al. (2024), who developed similar prompts to simulate student reasoning across skill levels on exam-style questions of varying difficulty. Their work informed our decision to incorporate reasoning about the ability of a student and the likelihood of error into our prompt design.

Our aim is not to claim these are optimal prompts or to exhaustively search for the best possible formulations. Instead, we adopt straightforward, representative prompting strategies aligned with popular practices to focus our investigation on whether such methods meaningfully promote grade-level alignment in model behavior. This may limit the scope of our findings, but it allows us to isolate and evaluate the effects of targeted prompting on grade-sensitive reasoning.

5 Experimental Setup

5.1 Task Setup

We design our experiments based on the framework described in Section 4. The evaluation is conducted in two phases:

1. **Problem-Solving:** LLMs answer questions in a standard problem solving setting, without specific instructions on how to mimic human behavior. Their performance is compared to that of average students in different grade levels.

2. **Grade-Level Mimicking:** LLMs are explicitly instructed to emulate an average student of a specific grade level and respond as such.

In both phases, we apply a Rasch model to assess performance. Each question is treated as an individual item j, and each LLM is treated as an in-distribution test-taker i. The binary response of LLM i to the question j is represented as $s_{ij} \in \{0, 1\}$, where $s_{ij} = 1$ indicates a correct answer.

LLM	Open Source?	Parameter Count	Fine-Tuned?	Benchmark Scores	
				GSM8K (%)	MMLU (%)
LLaMA2-13B (Touvron et al., 2023)	✓	13B	✗	28.7	54.8
LLaMA2-70B (Touvron et al., 2023)	✓	70B	✗	56.8	68.9
LLaMA3.1-8B (Touvron et al., 2023)	✓	8B	✗	84.5	73.0
LLaMA3.1-70B	✓	70B	✗	95.1	86.0
Mistral-7B (Jiang et al., 2023)	✓	7B	✗	52.1	60.1
Qwen2.5-7B (Yang et al., 2024)	✓	7B	✗	85.4	74.2
Qwen2.5-Math (Yang et al., 2024)	✓	7B	✓	91.6	67.8
GPT-3.5-Turbo (OpenAI, 2023)	✗	–	✗	57.1	70.0
o3-Mini (OpenAI, 2025)	✗	–	✗	89.9	85.2
SocraticLM (Liu et al., 2024)	✓	7B	✓	60.6	–
LearnLM-1.5-Pro (Modi and the LearnLM Team, 2024)	✗	–	✓	–	–

Table 1: List of LLMs evaluated in our study, along with key descriptors about each model, i.e., open source availability, parameter size, whether the model is fine-tuned (as opposed to pretrained or instruction-tuned), and scores on reasoning and comprehension benchmarks GSM8K and MMLU (we omit scores that have not been released publicly by the respective model's paper or technical report).

5.2 Models

We select a diverse set of 11 LLMs (see Table 1) to ensure broad coverage across access types (open vs. closed), model sizes and training paradigms (pretrained vs. domain-finetuned). Our goal is to capture a range of capabilities relevant to reasoning and comprehension, as reflected in benchmarks like GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2020). We include both general-purpose models and those finetuned to specific domains. GPT-3.5-Turbo is included based on Benedetto et al. (2024), who suggest that it can adapt responses to the levels of student ability instructed. SocraticLM and LearnLM-1.5-Pro are fine-tuned on pedagogical data; therefore, they might have more accurate insights into the performance of students at different grade levels.

5.3 Evaluation

Measuring Problem-Solving Correctness All problems in our dataset are multiple choice questions (MCQs), which simplifies the evaluation of correctness: *a model's response is considered correct if the selected option matches the correct answer provided with the dataset.* This binary distinction between correct and incorrect responses makes the data well-suited for dichotomous (i.e.,

the answer can only be correct or incorrect) Item Response Theory (IRT) models. We found that model responses vary in structure and require a unified follow-up prompt to extract the predicted choice from each model response (see Figure 13 in the Appendix).

Estimating Grade-Level Alignment To address both research questions, we estimate how closely a model's performance aligns with that of an average student at a given grade level. To pin the origin and unit of the Rasch ability scale during marginal maximum-likelihood estimation, we follow the standard convention of treating examinee abilities as standard-normal, $\theta \sim \mathcal{N}(0, 1)$. Although not theoretically necessary, Embretson and Reise (2000) note that this assumption is a reasonable way to identify the latent trait because it fixes the zero point and variance without constraining the shape of the data.

Therefore, the average student has an ability parameter of zero ($\theta_{\text{avg}} = 0$).

The estimated ability parameter θ_i for a model i can be interpreted in relation to this benchmark. *The closer θ_i is to zero, the more the model's performance aligns with that of the average student.* We can also express this alignment using the percentile rank, computed via the cumulative distribution function (CDF) of the standard normal distribution, denoted by $\Phi(\theta)$:

$$\text{Percentile Rank} = \Phi(\theta) \times 100 \tag{3}$$

A percentile rank of 50 corresponds to the average student. Higher percentile ranks indicate higher levels of ability relative to the population. We use percentile rank as our main metric to measure LLM or student ability, as it has a fixed range (0-100 and is centered at 50), which allows for easier comparison of LLM alignment across subjects and grade levels.

6 Results & Discussion

In Table 2, we report each LLM's percentile rank in grade-level mathematics and reading comprehension questions under three conditions: unenforced prompting (P_U), grade-enforced prompting (P_E) and their difference (Δ). We report P_U for the best prompt for each LLM per subject (see Table 5 in the Appendix for the best prompts for each setting) which maximizes grade alignment, i.e., when percentile rank is closest to 50 (average). `Avg. Deviation` records the mean absolute deviation from P = 50 for the corresponding prompt settings. The baseline `Random Choice` reports the percentile scores achieved with a randomized option chosen for each problem. This setup allows us to:

- address **RQ1** by comparing model percentiles to the student mean (50$^{\text{th}}$ percentile) across grades and subjects, and
- address **RQ2** by (i) quantifying the effect of grade enforcement on LLM performance (**RQ2.1**) and (ii) evaluating whether these shifts mirror human student response patterns (**RQ2.2**).

For further context, Table 4 presents the accuracy of each LLM under the unenforced condition.

6.1 RQ1: Alignment under Unenforced Prompting

We ask whether the unenforced problem solving prompt generates outputs that align with that of the average student in each grade (P_U).

Mathematics. Most models, especially those scoring well on GSM8K, e.g., `LLaMA3.1-70B`, `Qwen2.5-Math`, `o3-Mini`, and`SocraticLM`, achieve high percentiles in every grade, overshooting all benchmarks and showing no alignment with any specific grade. This is also reflected in the high average deviation of 40.5, 35.0, and 32.9 percentile points, respectively, from the optimal P=50 mark. In contrast, smaller models with relatively poorer benchmark performance, such as `LLaMA2-13B` and `Mistral-7B`, exhibit lower percentiles and show better alignment between grades.

Reading. Similar to mathematics, the models demonstrate high average percentile scores in reading for grades 4 and 8, proving unsuitable for faithful student mimicking. The models in our pool align better with grade 12, with relatively lower average P_U values. Fine-tuned models (not tuned for grade-alignment), e.g., `Qwen2.5-Math`, `SocraticLM` – `Qwen2.5-Math` further tuned on pedagogical data, have a poorer overall performance, resulting in better alignment across grades.

Across grades and subjects, all models score well above the `Random Choice` baseline. Without enforced instructions, LLMs rarely self-calibrate to grade difficulty. They overshoot when capacity is high and align only when under-powered or off-domain.

LLM	Mathematics									Reading								
	Question Grade 4			Question Grade 8			Question Grade 12			Question Grade 4			Question Grade 8			Question Grade 12		
	P_U	P_E	Δ	P_U	P_E	Δ	P_U	P_E	Δ	P_U	P_E	Δ	P_U	P_E	Δ	P_U	P_E	Δ
LLaMA2-13B	63.7	66.1	+2.4	52.6	50.8	-1.8	48.6	66.5	+17.9	99.7	95.5	-4.2	94.9	87.6	-7.3	80.9	58.7	-22.3
LLaMA2-70B	85.5	33.5	-52.0	23.9	31.6	+7.6	42.4	57.8	+15.4	88.6	79.8	-8.8	72.3	69.1	-3.2	71.6	58.7	-12.9
LLaMA3.1-8B	96.8	85.5	-11.3	85.2	60.0	-25.2	79.5	69.3	-10.2	99.7	77.9	-21.8	96.7	92.7	-4.0	86.7	83.9	-2.8
LLaMA3.1-70B	99.6	97.6	-2.0	98.9	98.0	-0.9	96.1	97.0	+0.9	99.9	99.9	0.0	96.7	92.7	-4.0	83.9	80.9	-3.0
Mistral-7B	63.7	58.9	-4.8	43.5	49.0	+5.4	63.7	57.8	-5.9	94.3	67.9	-26.4	84.8	69.1	-15.7	83.9	55.5	-28.4
Qwen2.5-7B	99.6	18.5	-81.2	99.3	22.5	-76.7	96.1	30.3	-65.8	98.2	5.2	-93.1	98.2	7.8	-90.4	77.9	30.2	-47.7
Qwen2.5-Math	99.8	70.7	-29.1	98.5	96.1	-2.4	97.8	97.8	0.0	72.0	69.9	-2.0	36.5	29.4	-7.1	43.4	43.4	0.0
GPT-3.5_Turbo	89.0	44.7	-44.3	70.8	11.7	-59.1	79.5	45.5	-34.0	99.7	61.8	-37.8	98.2	65.9	-32.3	68.3	43.4	-24.9
o3-Mini	98.9	98.3	-0.6	98.5	99.5	+1.0	95.1	99.3	+4.2	99.3	99.9	+0.6	99.1	98.2	-0.9	86.7	86.8	+0.1
SocraticLM	99.3	96.8	-2.6	99.7	99.3	-0.5	97.0	98.4	+1.4	59.8	63.9	+4.1	34.0	39.1	+5.0	32.6	49.3	+16.7
LearnLM-1.5-Pro	99.8	75.3	-24.6	99.7	93.6	-6.2	98.9	98.4	-0.5	99.9	24.5	-75.4	94.9	53.3	-41.6	65.1	58.7	-6.4
Avg. Deviation	40.5	27.5	-13.0	35.0	30.2	-4.8	32.9	28.8	-4.2	41.9	30.6	-11.3	37.8	27.5	-10.3	25.4	15.2	-10.2
Random Choice	6.1			1.4			6.7			4.12			4			0.9		

Table 2: LLM percentile scores on grade-level questions from mathematics and reading without grade enforcement (P_U – shaded blue), with grade enforcement (P_E – shaded green), and their difference (Δ = P_U - P_E – shaded yellow). Darker hues for P_U and P_E denote closer alignment to the average score of 50 and larger absolute change in Δ. **Boldface** highlights the best model (i.e., closest to 50) in each setting. `Avg. Deviation` records the mean absolute deviation from P=50 for corresponding prompt settings. The `Random Choice` baseline reports the percentile scores attained with a randomized option chosen for each problem.

6.2 RQ2.1: Effect of Grade-Level Prompts

We test whether prompting a model to "think like an average grade g student" changes its performance (Δ in Table 2), regardless of the resultant alignment.

Drops: Overall, we note that upon prompting models to mimic the average student of grades 4, 8, or 12, percentile scores generally drop (see Avg. Deviation Δ in Table 2), with a greater drop for a lower target grade. `Qwen2.5-7B` records the highest drop of 93.1 percentile points for reading grade 4 as well as the greatest average drop.

Gains: We observe that grade-specific prompting can also increase model performance. For example, several settings with `LLaMA2-13B/70B` for mathematics and all grade settings with `SocraticLM` for reading result in higher P_E than P_U.

Stable: Some models, such as `LLaMA3.1-70B` and o3-Mini for subjects and `SocraticLM` for mathematics, show little to no change between their values of P_U and P_E values, despite their respective P_U values having a high deviation from the target P=50.

Prompt Strength: Among the three grade enforcement prompts, the most detailed GRADEENFORCEDFULLCOT prompt (with explicit instruction to consider the probability that an average student of the target grade will get the given problem right) causes the largest changes (Figure 3a). This shows that grade-level cues can markedly increase or lower scores depending on the model and prompt strength, although a few models remain robust.

6.3 RQ2.2: Alignment Under Enforced Prompts

We investigate whether grade-specific prompts move the model performance closer to the average student (ideal $P = 50$). We find that the results are spread across the following categories:

(1) Aligned P_U and aligned P_E: Some models that are close to the 50th percentile without grade-specific prompting maintain good alignment after prompting (for example, `LLaMA2-13B / 70B` for mathematics and `SocraticLM` for reading). These models can act as "proxy students" out of the box for particular pairs of subjects' grades.

(2) Misaligned P_U and misaligned P_E: Other models' percentile scores can range far above or below the median despite grade-specific prompting (for example, P_Us and P_Es for `o3-Mini` across subjects and grades stay far above 50). We did not observe any model that consistently scored below the median percentile.

(3) Misaligned P_U and aligned P_E: In some cases, prompting can help induce grade alignment (P_E) when unenforced alignment is poor (P_U). For example, `Mistral-7B`'s percentile range on reading problems moves from $P_U \in [83.9 - 94.3]$ to P_E=$[55.5 - 69.1]$; `GPT-3.5-Turbo` shows similar gains in most tasks. Such cases demonstrate the desired effect of grade-specific prompting.

(4) Aligned P_U and misaligned P_E: In contrast, grade-specific prompting can cause models to overshoot. `Qwen2.5-7B` in grade 4 reading drops from 98.2 to 5.2 (Δ=-93.1), overshooting the target.

Prompt design matters. Figure 3b shows that the GRADEENFORCEDFULLCOT template changes scores the most. However, it is not always the most optimal prompt setting to achieve better grade alignment (lower percentile deviation from 50).

Fine-tuned models. Pedagogically tuned models (`LearnLM-1.5-Pro`, and `SocraticLM`) are not better aligned than general LLMs (such as `Mistral-7B`), with or without prompts, indicating that faithful grade-level emulation probably needs explicit alignment objectives.

Thus, grade alignment is model-prompt specific; no single prompt works everywhere. Reliable grade-level emulation will require tailored prompting that does not ensure generalization to other grades or subjects.

6.4 Guidelines for Selecting Viable LLM "Proxy Students"

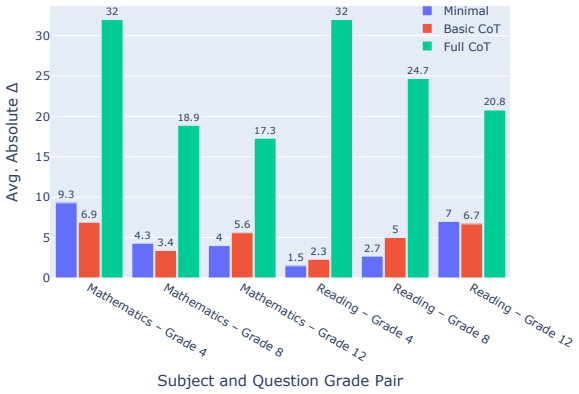
Our experiments confirm that no single model–prompt pair reliably matches average student performance in every grade and subject. Before an LLM can stand in for real students, e.g., to trial new test items or train a model for an ITS, it should pass the following baseline checks:

1. **Grade alignment.** The model's ability estimate (θ_n) in a representative item set must fall within the normative band of the grade: core average ± 1 logit (percentile: 15.9 to 84.1), extended ± 1.5, outlier $\geq \pm 2$ (Bond and Fox, 2015). Models such as `GPT-3.5-Turbo` stayed in the core range with appropriate prompts for most grades.

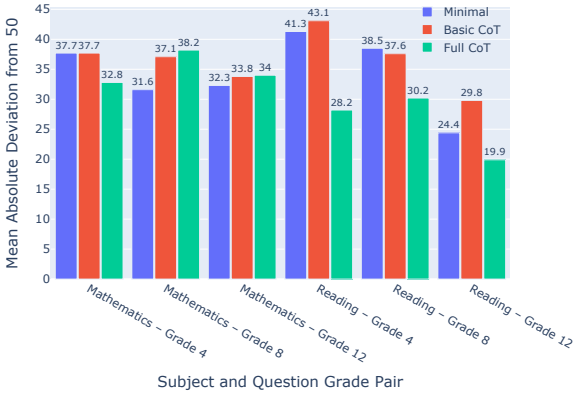
2. **Developmental ordering.** Ability should rise monotonically with grade, mirroring trends in NAEP reading (217, 262, and 285 according to NAEP's own cumulative scoring scale for grades 4, 8, 12 respectively; National Center for Education Statistics, 2022). Several pairs violated this; e.g., `Mistral-7B`'s P_U was 38.3, 17.0, 40.5 for the same grades.

3. **Prompt stability.** Grade-enforcing prompts can improve or harm performance. An unenforced prompt should be used if the model is already aligned; otherwise, one should verify that enforcement is equally accurate across all grades.

These criteria are necessary, but not sufficient. We believe that more accurate guidelines for faithful student mimicking will emerge with richer evaluation datasets.



(a) On Δ



(b) On grade-alignment. We report mean absolute deviation from an average grade-level student's percentile score (i.e., 50). Thus, greater the deviation, poorer the average alignment.

Figure 3: Impact of grade-enforcing prompting

7 Conclusion

In this paper, we investigate whether LLMs' regular problem-solving performance aligns with that of an average student of a given grade, and whether

explicit prompting to act like the average student makes a difference and improves this alignment. We conduct a thorough analysis of 11 diverse models on mathematics and reading questions from K-12 grades 4, 8, and 12 sourced from the NAEP database. Our IRT-based analysis reveals that in the regular (unenforced) setting, stronger models score far better than the average students of any grade and weaker models may align well incidentally. Though explicit (grade-enforced) prompting causes a change in model performance, the alignment with the desired grade-level average varies substantially across model and prompt combinations, with no single model-prompt pair producing average performance across grades or subjects. We provide a set of necessary guidelines to select viable student-proxies for future work and highlight the need for dedicated model finetuning for faithful grade adherence.

Limitations

While the results of our experiments lead to certain conclusions and provide us with novel insights, we acknowledge that these are necessarily limited in a number of ways.

Limited number of samples and subjects considered: Getting access to publicly available student answering data is challenging. The NAEP (National Center for Education Statistics, 2022) database offers a valuable resource in that regard. However, the database is not naturally designed to provide data for performing analysis over automated models at scale, therefore, the available subjects and the number of questions in the collected dataset are limited.

Text-based questions only: In this study, we have restricted our analysis to text-only questions, omitting questions that involve visual interpretation. We admit that this is not completely faithful to student assessments, as visual cues may also elicit key reasoning abilities. We plan to expand our study to more modalities in the future.

MCQ format: Evaluation of LLM responses is a key challenge, especially for free-form answering style. To mitigate this challenge, in this work, we focus on MCQ-type questions only. This also makes modeling the items within the IRT framework easier. As models vary in their response structure, we find that simple rule-based extraction is not reliable enough, and we have to use a follow-up prompt to extract the final option selected by the model. We plan to develop more robust evaluation strategies to allow for more varied question types in the future.

No data for cross-grade performance used: A key point to note is that NAEP only reports the performance of students at a particular grade level on questions from the same grade. Though this is adequate for assessing student learning trends, for determining cross-grade viability of proxy students, we would require real students' performance on questions from different grades.

Use of prompting methods only: We focus our study solely on prompt-based methods to enforce grade-level alignment, as this is one of the most accessible ways in which models are used in this context, as demonstrated by previous work. A more in-depth analysis is needed to assess whether in-context learning and finetuning strategies can also play a role in improving the quality of proxy-students, in addition to appropriately sized student demonstration data for tuning. We also highlight that prompt engineering (i.e., designing the most optimal prompts for the models) was outside the scope of this study, and the prompts that we used are inspired by previous work in this domain.

Experiments with specific models: Last but not least, we acknowledge that our findings apply to a specific set of models considered in this study. We highlight that our choice was motivated by considerations around the diversity of the model pool.

Ethical Considerations

This study relies exclusively on cumulative, de-identified statistics drawn from student response data supplied by the National Assessment of Educational Progress (NAEP). No record contains direct or indirect identifiers, and at no stage were individual-level student data accessed, stored, or analyzed. All analytic procedures conformed to the NAEP Data Confidentiality and Disclosure Policy as well as the privacy protections required under the Family Educational Rights and Privacy Act (FERPA). Consequently, the research poses no risk to the privacy or well-being of individual students.

References

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

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. 2015. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3 edition. Routledge, New York.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Susan E. Embretson and Steven P. Reise. 2000. *Item Response Theory for Psychologists*. Multivariate Applications Series. Lawrence Erlbaum Associates, Mahwah, NJ.

Michael Grohs, Luka Abb, Nourhan Elsayed, and Jana-Rebecca Rehse. 2024. Large language models can accomplish business process management tasks. In *Proceedings of the International Conference on Business Process Management*. Extended version available as arXiv:2307.09923.

Joy He-Yueya, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W. Domingue, Emma Brunskill, and Noah D. Goodman. 2024. Psychometric alignment: Capturing human knowledge distributions via language models. *Preprint*, arXiv:2407.15645.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *arXiv preprint*.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. Socraticlm: Exploring socratic personalized teaching with large language models. In *Advances in Neural Information Processing Systems (NeurIPS) 2024*.

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56:1028–1052.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Noboru Matsuda, Dan Lv, and Guoliang Zheng. 2023. Teaching how to teach promotes learning by teaching. *International Journal of Artificial Intelligence in Education*, 33(3):720–751.

Abhinit Modi and the LearnLM Team. 2024. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*.

Ethan R. Mollick, Lilach Mollick, Natalie Bach, L. J. Ciccarelli, Ben Przystanski, and Daniel Ravipinto. 2024. Ai agents and education: Simulated practice at scale. *arXiv preprint*.

National Center for Education Statistics. 2022. The nation's report card: 2022 naep reading and mathematics assessments. `https://nces.ed.gov/nationsreportcard/`. Accessed: 2025-04-20.

OpenAI. 2023. GPT-3.5-Turbo [large language model]. `https://platform.openai.com/docs/models/gpt-3-5-turbo`. Accessed: 2025-04-25.

OpenAI. 2025. OpenAI o3-mini [large language model]. `https://platform.openai.com/docs/models/o3-mini`. Accessed: 2025-04-25.

Georg Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen. Reprinted by University of Chicago Press (1980) and MESA Press (1992).

Glen Smith, Adit Gupta, and Christopher J. MacLellan. 2024. Apprentice tutor builder: A platform for users to create and personalize intelligent tutors. *arXiv preprint*.

Shashank Sonkar, Xinghe Chen, Naiming Liu, Richard G Baraniuk, and Mrinmaya Sachan. 2024. Llm-based cognitive models of students with misconceptions. *arXiv preprint arXiv:2410.12294*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

UNESCO. 2023. Global education monitoring report 2023: Technology in education. https://unesdoc.unesco.org/ark:/48223/pf0000385723.

U.S. Department of Education. 2023. Artificial intelligence and the future of teaching and learning: Insights and recommendations. https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Joel Williams. 2003. *The Skills for Life survey: A national needs and impact survey of literacy, numeracy and ICT skills*. 490. The Stationery Office.

Beverly Woolf, Ivon Arroyo, and 1 others. 2013. Intelligent tutoring systems by and for the developing world: A review of trends and opportunities. *International Journal of Artificial Intelligence in Education*, 24(3):331–367.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason 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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2205, Singapore. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

A Dataset Details

Table 3 presents the distribution of the questions extracted from NAEP by subject and grade levels. Figure 4 shows an example NAEP question from grade 12 reading. We use data that is publicly available on the NAEP website.

Subject	Grade	Number of Questions	Percentage Share
Mathematics	4	82	32.93
	8	106	42.57
	12	61	24.50
	Total	249	100.00
Reading	4	101	42.08
	8	72	30.00
	12	67	27.92
	Total	240	100.00

Table 3: Dataset statistics: Number of Questions across Subjects and Grade-Levels

B Prompt Templates

Figures 5 to 12 show the different solution generation prompts for mathematics and reading. Figure 13 shows the prompt used to extract the final option from the generated solution.

C Querying Setup

All models were queried with the following hyperparameters: `temperature`=0, `top_p`=0.95, and `max_tokens`=2048. `LLaMA3.1` models were queried using the Google Cloud (Vertex) API, `o3-Mini` was queried using the OpenAI API, and `LearnLM-1.5-Pro` was queried using Google's AI Studio API. All other models were imported from HunggingFace and queried locally using vLLM on a single NVIDIA A100 GPU. Each round of querying took less than one hour.

D Model Ability (θ_n) Estimation Algorithm

Algorithm 1 captures the steps required to fit the Rasch model as described in §4.

E Analyses Details

Table 4 lists LLMs' accuracy on mathematics and reading problems from different grade levels with the unenforced prompt setting. Table 5 records the best prompt out of the four possible settings, depending on the closeness of the corresponding percentile values to 50.

Algorithm 1 Estimating LLM Ability and Percentile Rank Using the Rasch Model

Require: $p = \{p_j\}_{j=1}^{I}$ ▷ Proportion of correct responses for item j across students
Require: $s = \{s_{ij}\}_{i=1,j=1}^{M,I}$ ▷ Binary correctness matrix: LLM i's response to item j
Ensure: $\theta = \{\theta_i\}_{i=1}^{M}$ ▷ Estimated ability (logit scale) for each LLM
Ensure: $\pi = \{\pi_i\}_{i=1}^{M}$ ▷ Percentile rank of each LLM

Step 1: Estimate item difficulties using student response proportions
for $j = 1$ to I **do**
    $b_j \leftarrow \log\left(\frac{1-p_j}{p_j}\right)$ ▷ Item difficulty via inverse of the Rasch probability function
end for
Step 2: Estimate LLM abilities via maximum likelihood using the Rasch model
for $i = 1$ to M **do**
    Define likelihood function:

$$\mathcal{L}(\theta_i) = \sum_{j=1}^{I} s_{ij} \cdot log\left(\frac{1}{1+e^{-(\theta_i - b_j)}}\right) + (1 - s_{ij}) \cdot log\left(1 - \frac{1}{1+e^{-(\theta_i - b_j)}}\right)$$

    Estimate $\theta_i = \arg\max_\theta \mathcal{L}(\theta)$ ▷ MLE for the Rasch model
end for
Step 3: Compute LLM percentile ranks w.r.t. student ability distribution
Let $\Phi(\theta)$ be the cumulative distribution function (CDF) of student abilities
for $i = 1$ to M **do**
    $\pi_i \leftarrow \Phi(\theta_i) \times 100)$ ▷ Percentile rank of LLM i
end for

LLM	Mathematics			Reading		
	4	8	12	4	8	12
LLaMA2-13B	78.05	43.40	41.67	85.15	80.56	77.61
LLaMA2-70B	65.85	59.43	45.00	96.04	91.67	82.09
LLaMA3.1-8B	87.80	77.36	63.33	96.04	93.06	85.07
LLaMA3.1-70B	93.90	91.51	80.00	98.02	93.06	83.58
Mistral-7B	65.85	54.72	53.33	89.11	86.11	83.58
Qwen2.5-7B	93.90	92.45	80.00	93.07	94.44	80.60
Qwen2.5-Math	95.12	90.57	83.33	76.24	63.89	64.18
GPT-3.5_Turbo	80.49	68.87	63.33	96.04	94.44	76.12
o3-Mini	91.46	90.57	78.33	95.05	95.83	85.07
SocraticLM	92.68	94.34	81.67	70.30	62.50	58.21
LearnLM-1.5-Pro	95.12	94.34	86.67	97.03	91.67	74.63
Average	78.69	72.13	64.06	83.01	79.60	71.90

Table 4: LLM accuracy scores (i.e., accuracy in solving the tasks) for different grade levels in mathematics and reading under the unenforced prompt setting.

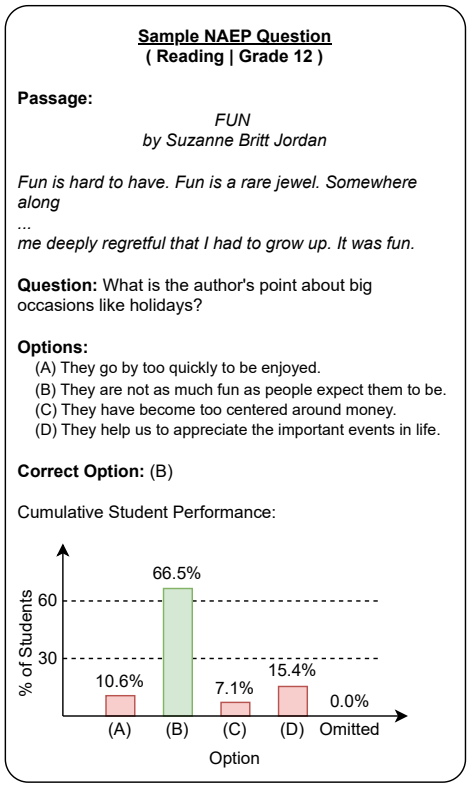


Figure 4: Sample NAEP question from grade 12 reading.

Math - Unenforced Minimal Prompt

```
Answer the given question to
the best of your ability.

Question: <<QUESTION>>
```

Figure 5: UNENFORCED Prompt Template for Mathematics

Math - Enforced Minimal Prompt

```
You are an average grade
<<GRADE>> student.
Answer the given question.

Question: <<QUESTION>>
```

Figure 6: GRADEENFORCEDMINIMAL Prompt Template for Mathematics

Math - Enforced Basic-CoT Prompt

```
You are an average grade
<<GRADE>> student.
Answer the given question.
Your reasoning,
knowledge-level, and
tendency to make different
kinds of mistakes should
reflect your specified grade
level.

Question: <<QUESTION>>
```

Figure 7: GRADEENFORCEDBASICCOT Prompt Template for Mathematics

Math - Enforced Full-CoT Prompt

```
You are an average grade
<<GRADE>> student.
Answer the given question.
First reflect upon whether
an average grade <<GRADE>>
student is likely to answer
the question correctly. If
so, answer correctly. Else,
provide the most likely
incorrect answer an average
grade <<GRADE>> student
would pick.

Question: <<QUESTION>>
```

Figure 8: GRADEENFORCEDFULLCOT Prompt Template for Mathematics

Reading - Unenforced Minimal Prompt

```
Read the given passage and answer
the question at the end to the
best of your ability.

---- Reading Passage Begins ----

<<READING_PASSAGE>>

---- Reading Passage Ends ----

Question: <<QUESTION>>
```

Figure 9: UNENFORCED Prompt Template for Reading

Reading - Enforced Minimal Prompt

```
You are an average grade <<GRADE>>
student.
Read the given passage and answer the
question at the end.

---- Reading Passage Begins ----

<<READING_PASSAGE>>

---- Reading Passage Ends ----

Question: <<QUESTION>>
```

Figure 10: GRADEENFORCEDMINIMAL Prompt Template for Reading

Reading - Enforced Basic-CoT Prompt

```
You are an average grade <<GRADE>>
student.
Read the given passage and answer the
question at the end.
Your reasoning, knowledge-level, and
tendency to make different kinds of
mistakes should reflect your specified
grade level.

---- Reading Passage Begins ----

<<READING_PASSAGE>>

---- Reading Passage Ends ----

Question: <<QUESTION>>
```

Figure 11: GRADEENFORCEDBASICCOT Prompt Template for Reading

LLM	Best Prompt for Mathematics	Best Prompt for Reading
LLaMA2-13B	GRADEENFORCEDMINIMAL	GRADEENFORCEDMINIMAL
LLaMA2-70B	GRADEENFORCEDBASICCOT	GRADEENFORCEDFULLCOT
LLaMA3.1-8B	GRADEENFORCEDMINIMAL	GRADEENFORCEDFULLCOT
LLaMA3.1-70B	GRADEENFORCEDBASICCOT	GRADEENFORCEDMINIMAL
Mistral-7B	GRADEENFORCEDMINIMAL	GRADEENFORCEDFULLCOT
Qwen2.5-7B	GRADEENFORCEDFULLCOT	GRADEENFORCEDFULLCOT
Qwen2.5-Math	GRADEENFORCEDFULLCOT	GRADEENFORCEDMINIMAL
GPT-3.5-Turbo	GRADEENFORCEDFULLCOT	GRADEENFORCEDFULLCOT
o3-Mini	GRADEENFORCEDFULLCOT	GRADEENFORCEDFULLCOT
SocraticLM	GRADEENFORCEDFULLCOT	GRADEENFORCEDFULLCOT
LearnLM-1.5-Pro	GRADEENFORCEDFULLCOT	GRADEENFORCEDFULLCOT

Table 5: Best prompts for LLM in each subject. We pick the best prompt based on the closest average percentile rank to 50, i.e., the desired average performance.

Reading - Enforced Full-CoT Prompt

```
You are an average grade
<<GRADE>> You are an average
grade <<GRADE>> student.
Read the given passage and
answer the question at the end.
First reflect upon whether an
average grade <<GRADE>> student
is likely to answer the question
correctly. If so, answer
correctly. Else, provide the
most likely incorrect answer an
average grade <<GRADE>> student
would pick.

---- Reading Passage Begins ----

<<READING_PASSAGE>>

---- Reading Passage Ends ----

Question: <<QUESTION>>
```

Figure 12: GRADEENFORCEDFULLCOT Prompt Template for Reading

Option Extraction Prompt

```
You are given a response to a MCQ
problem.

===== RESPONSE START =====
<MODEL_RESPONSE>
===== RESPONSE END =====

What is the final answer opted by the
response?

The answer can only be one of the
objective choices: (A), (B), (C), (D),
(E), etc.

You are not to solve any part of the
underlying problem yourself.

The final answer you return should be
based solely on the given response.

Note that it is possible that the
response suggests that one option is
correct, however it provides another
option as the final answer. You are
expected to return the latter in such
cases.

Only reply with the final answer.
```

Figure 13: Option Extraction Prompt