20th BEA Workshop

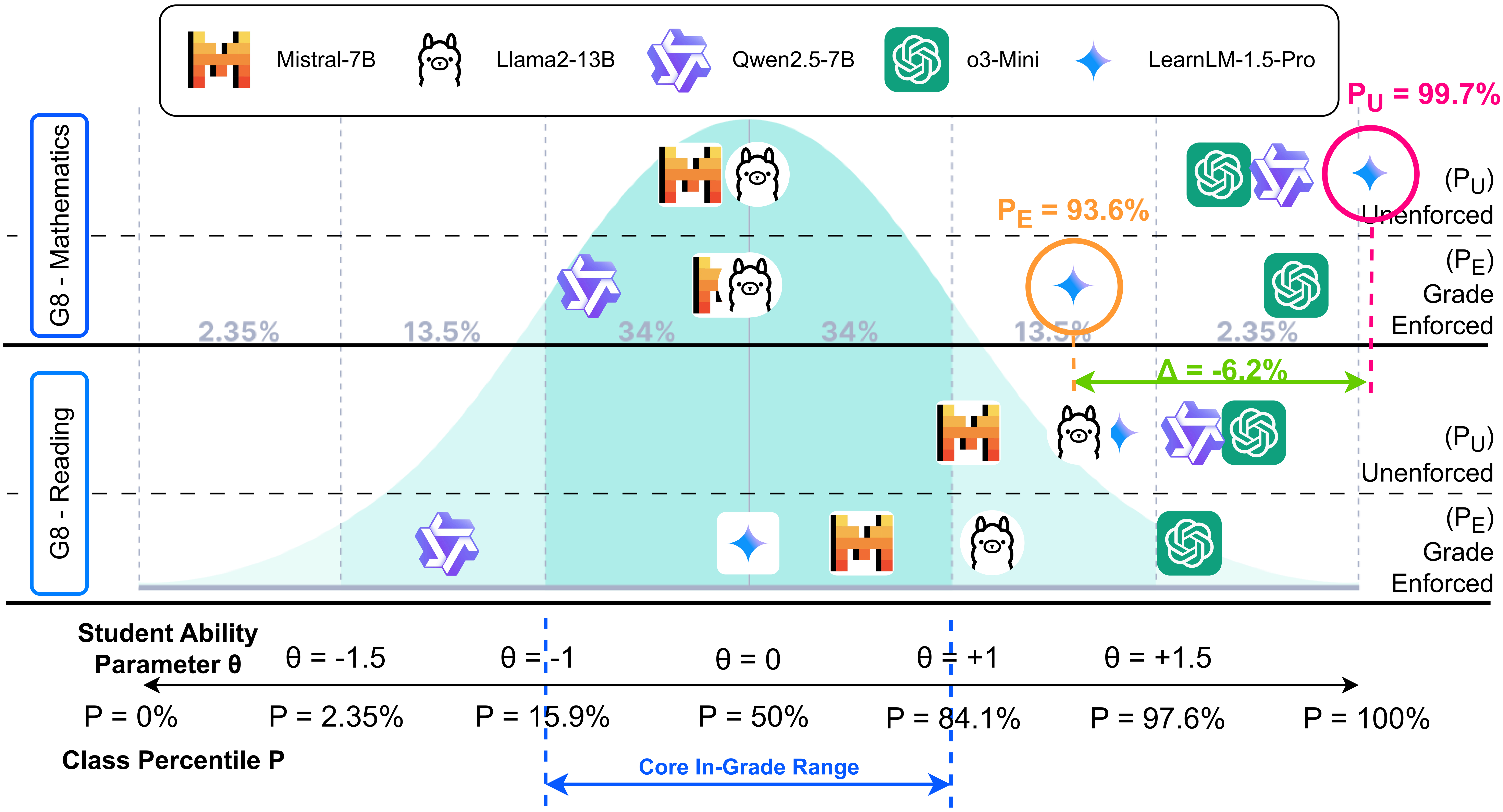
ACL 2025 VIENNA JULY 27 - AUGUST 1

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 (MBZUAI), Abu Dhabi, UAE

Mohamed bin Zayed University of Artificial Intelligence



Problem Statement

Can LLMs faithfully emulate the response patterns of students **of a specific grade?**

Motivation

Access to real student data is challenging. Using LLM-based proxy students can aid in

+ Evaluating and developing tutoring systems
+ Pilot testing new assessments.

Method

We compare LLMs' performance on MCQ test questions from the **National Assessment of Educational Progress (NAEP)** [1] with students' overall response behavior across grades.

Using a simple **Rasch model** from Item Response Theory (IRT), we determine each model's ability parameter (θ). Mapped to class percentile, the ability estimate allows us to compare an LLM's performance with that of the **average student of a given grade**.

An LLM with ability close to that of the average student (θ = 0 or P = 50%) is deemed better aligned [2].

We conduct our study in two settings:
1. **Unenforced:** Using a regular problem-solving prompt to measure native ability.

2. **Grade Enforced:** Explicitly instructing the model to *"act"* like the average student of a given grade. We test <u>three prompts</u> with successively greater descriptiveness.

I - Unenforced (P_U)

Mathematics: Strong models overshoot the average (no alignment with any grade). Weaker models (e.g., Mistral-7B) show better alignment.

Reading: LLMs generally struggle with reading problems and thus, show better alignment with grade 12.

Best Practices

Grade Alignment: Must fall within normative grade bands -- Core (15.9% - 84.1%).

Developmental Ordering: For cross-grade proxies, ensure proper performance gradation, i.e., $P_4 <= P_8 <= P_{12}$

Prompt Stability: Grade enforced performance is volatile and case-based. Stick to unenforced querying unless necessary.

References

[1] National Center for Education Statistics. 2022. The nation's report card: 2022 naep reading and mathematics assessments. https://nces.ed.gov/nationsreportcard/.

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

Data & Code | Presentation | arXiv Paper

II - Grade Enforced (P_E)

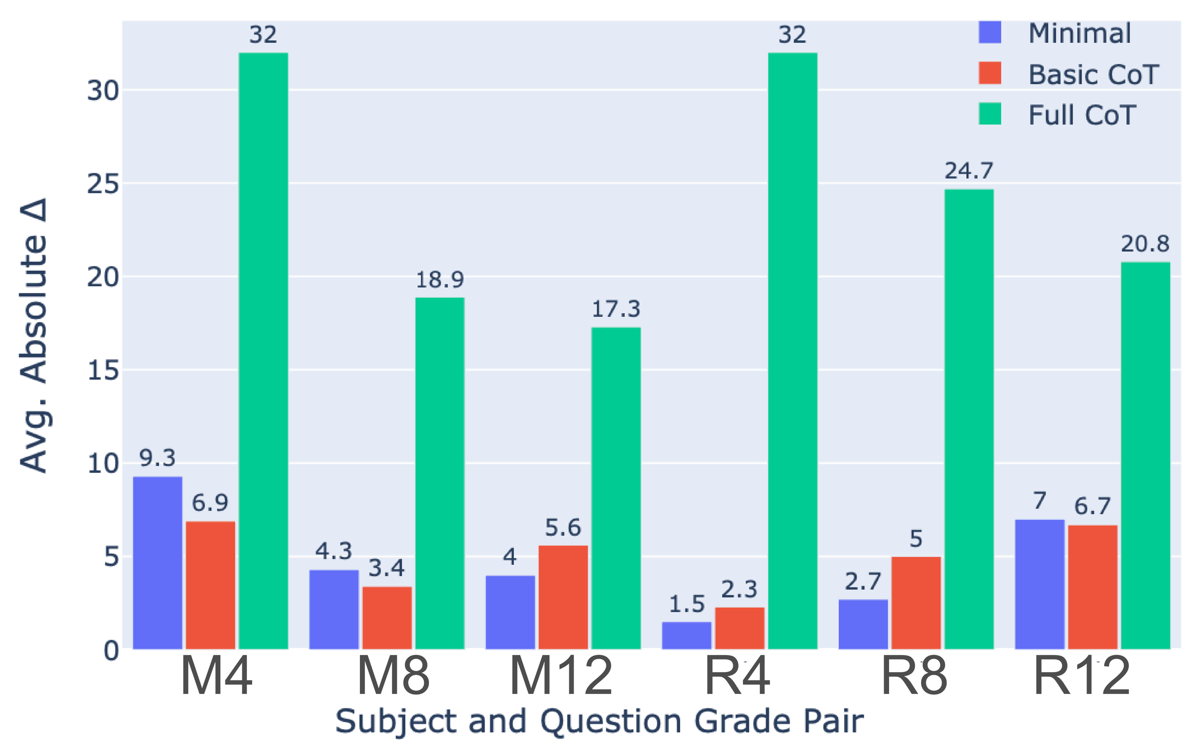
When testing grade enforcing prompts, we evaluate two aspects:

II - (a) Change ($\Delta = P_E - P_U$)

Drops: Greater drops for lower target grades.

Gains: LLM scores can improve when asked to mimic higher grade.

Prompt Strength: More descriptive prompting leads to greater shifts.



II - (b) Alignment

Alignment is possible; But, **no single prompt or model combination works always.**

Prompt Strength ≠ Accuracy: greater shifts do not always furnish better alignment.

No significant benefit from fine-tuning. Models tuned on math or pedagogical data do not exhibit better alignment than others.