NCSA 2026

The “Too Good” Problem

If AI struggles to simulate a struggling student, what does that mean for item difficulty prediction?

Sarah Quesen, PhD MPH
Senior Director of Assessment
WestEd

As LLMs keep getting better, do we have a “too good” problem?

- Research suggests that LLMs with weaker ability produced *better* simulations than stronger models.
- High-accuracy models seem unable to convincingly produce poor responses, which compresses the lower end of the simulated ability distribution.

Acquaye, C., Huang, Y. T., Carpuat, M., & Rudinger, R. (2026). arXiv. https://doi.org/10.48550/arXiv.2601.09953

The Problem We Set Out to Explore

Standard setting panels may benefit from student response examples before operational scoring is complete.

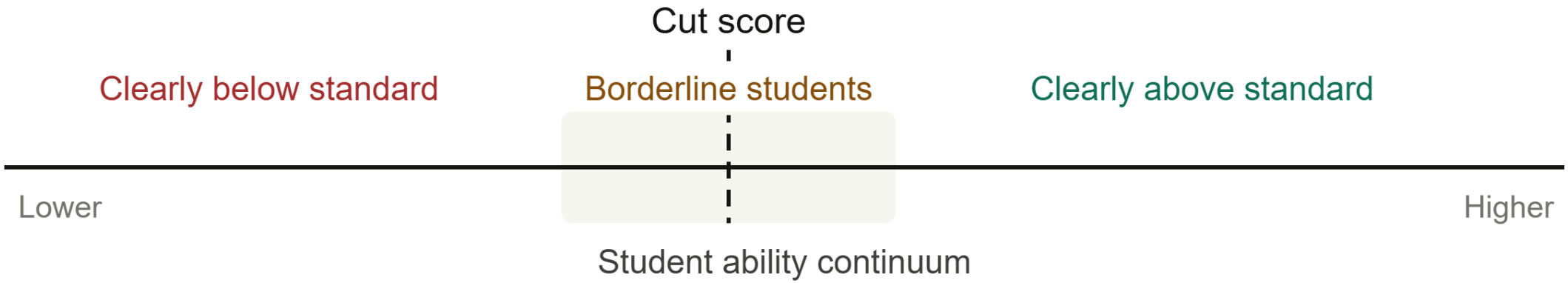
Recent work uses LLMs to simulate student responses for item calibration and difficulty prediction (e.g., Scarlatos et al., 2025; Acquaye et al., 2026; Srivatsa et al., 2025; Benedetto et al., 2024).

No prior work generates responses as panelist training materials.

RQ: *Can LLMs generate written responses at specified IRT ability levels that vary in quality consistent with the expected ability location?*

The Challenge

- Standard setting requires a focus on a very narrow range of the ability scale, where students are JUST beyond the threshold into the next performance level.



The borderline student sits right at the cut: just barely meeting the standard.

Study Data

Five NAEP Grade 8 Reading Items

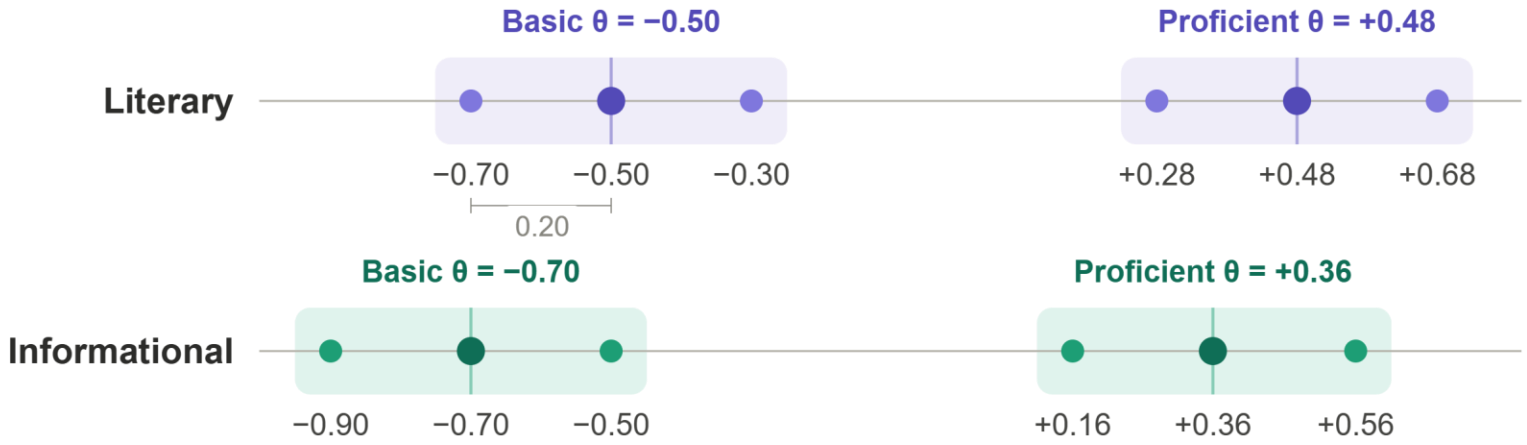
Item	Passage	What students do	Max Points
R079504	literary	Describe character	2
R079507	literary	Character change	3
R079508	literary	Interpret statement	1*
R079510	literary	Most important character	2
R069707	info	Identify evidence types	3

*R079508 was later dropped because the model could not generate meaningful variation on a dichotomous item.

- Items are publicly released with sample responses, published scoring rubrics, and IRT parameters.

Target Ability

Where on the scale did we attempt to generate responses?



- We solved for the approximate theta cuts based on the NAEP scale transformation.
- We generated 10 student responses × 5 items × 6 theta levels for each condition.

Study Methods: Two Large Language Models

Fine-tuned model

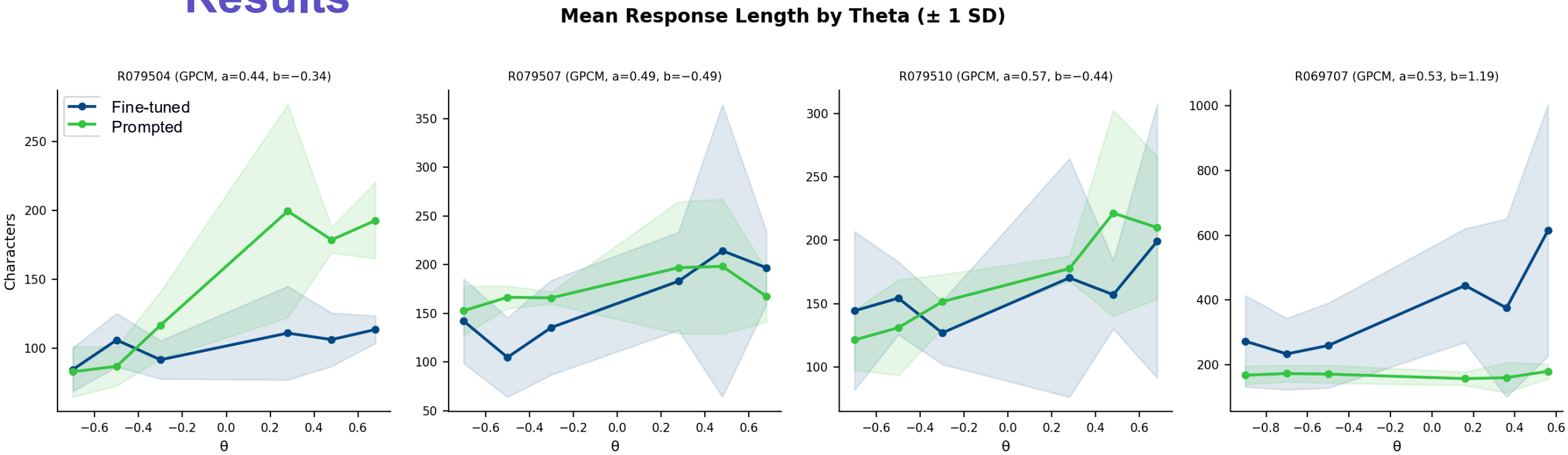
- Open-source AI model (Qwen3-4B), trained specifically for this task
- Training data: >30 student responses, each mapped to an ability level estimate using item parameters
- After training, the model generates new responses at any requested point on the ability scale

Prompted model

- General-purpose AI model (Gemini 2.5 Flash), no training
- Each prompt includes the scoring rubric, scored student exemplars with ability estimates, and the expected score distribution at the target ability level (few-shot prompting)

WestEd

Results



- The first question is whether the models respond to the theta signal at all. Response length is a simple way to see this.
- The prompt-only model for the hardest item (R069707) didn't move the needle. Also, this was the only informational item type in the study.

Results. *Describe what kind of person the merchant is. Give one detail from the story to support your answer.*

Fine-tuned sample responses

Basic cut region

"The merchant is greedy because he asked for 10,000 akches for 5 eggs."

Proficient cut region

"The merchant is a kind and honest person because he never forgot his humble beginnings or the money he owed the innkeeper."

Prompted sample responses

Basic cut region

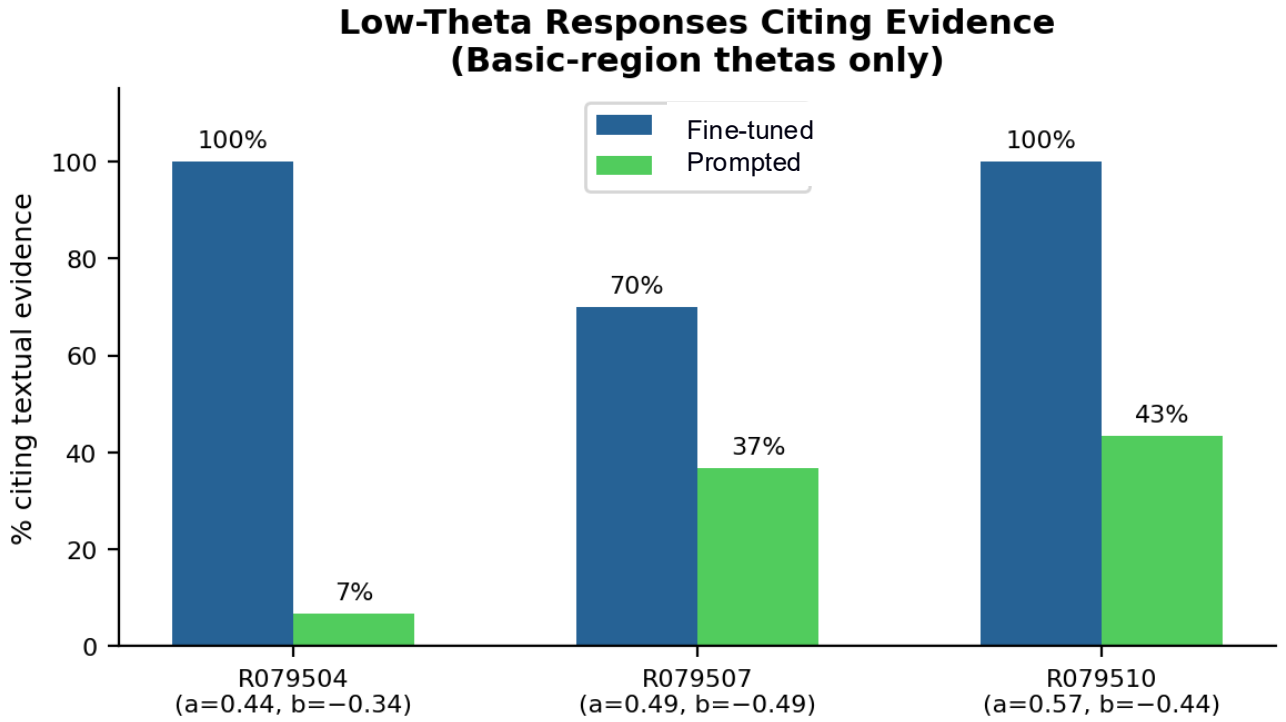
"The merchant is a honest man. He remembered to pay back the innkeeper after a long time."

Proficient cut region

"The merchant is a really honest and responsible person. He showed this because even after many years and becoming a rich sea merchant, he still remembered the five boiled eggs he ate when he was poor."

WestEd

The "Too Good" Problem



But the fine-tuned model produces evidence-citing responses 70 to 100% of the time at thetas where students should show limited ability.

What This Means

- Neither method is ready for operational use, but we show proof of concept.
- Both methods perform poorly at realistic low-ability responses, but in somewhat different ways.
- **Human review isn't optional.**
- Any testing program with a secure LLM chatbot, items, rubrics, and IRT parameters could use a prompted model.
    - Worth further research!
      Different approaches to prompting could improve results.

Relating This Study Back to Difficulty Prediction

Model cannot produce convincingly low-ability work.

Generated low-ability responses score too high.

Calibrate on those scores and the item looks easier than it is.

Difficulty is biased low, especially for difficult open-ended items.

Questions for States to Consider

- Is there evidence that the model can meet expectations at the tails of the distribution?
- Is there evidence from human expert panels scoring the responses vs model-based predictions?

Connect with WestEd

WestEd.org/subscribe

Questions?

References

Benedetto, L., et al. (2024). Using LLMs to simulate students' responses to exam questions. Findings of the ACL.

Acquaye, C., Huang, Y. T., Carpuat, M., & Rudinger, R. (2026). Take out your calculators: Estimating the real difficulty of question items with LLM student simulations. arXiv:2601.09953.

Scarlatos, A., Fernandez, N., Ormerod, C., Lottridge, S., & Lan, A. (2025). SMART: Simulated students aligned with item response theory for question difficulty prediction. Proceedings of EMNLP 2025.

Srivatsa, K. A., et al. (2025). Can LLMs reliably simulate real students' abilities in mathematics and reading comprehension? Proceedings of the BEA Workshop, ACL 2025.

NAEP 2017 Grade 8 Reading: Released items, IRT parameters, and scoring guides. nationsreportcard.gov.