ASSISTments

PUBLISHED RESEARCH

FoundationalASSIST

The first dataset that lets AI systems actually understand how students learn

DATE

September 2026

AUTHORS

Cristina Heffernan

The Gap We Set Out To Fill

Imagine trying to teach a child mathematics, but every time a student answers a question, all you are told is whether they are “right” or “wrong.” You never see what the student actually wrote, which mistake they made, or what misunderstanding might be driving the error. A tutor with such limited information is being asked to diagnose learning without the evidence needed to do so.

For years, educational researchers have faced a similar limitation. Many of the most widely used learning datasets, including the ASSISTments datasets from Worcester Polytechnic Institute, captured millions of student interactions primarily as a student ID, a problem ID, and a binary outcome: correct or incorrect. While these datasets are valuable for certain traditional models, richer, more diverse data formats, such as natural language text, can better leverage the capabilities of large language models (LLMs) in educational research and practice.

An LLM confronting these outdated datasets sees sequences of identifiers and binary outcomes. It cannot fully leverage its language understanding because there is no language to understand.

What FoundationalASSIST Provides

FoundationalASSIST is the first English-language educational dataset designed for the age of AI and foundation models. Built from real student activity on the ASSISTments platform between 2019 and 2024, it has the information needed to support research.

1.7M
student interactions

5,000
unique students

3,395
math problems with full text

The dataset preserves the complete text of each math problem, not just an identifier, so an AI can read and reason about it. It records what each student actually wrote as their answer (whether that's a number, a fraction, an expression, or a multiple-choice selection). And it tracks which specific wrong answer each student answered, allowing the LLM to investigate the misconception behind every error.

All 3,395 problems are drawn from the Illustrative Mathematics curriculum for grades 6–8 and aligned to the Common Core State Standards. Every problem is tagged with a grade level and a standard, giving researchers a structured view of how middle school math is actually taught in American classrooms.

Why This Changes What Is Possible

Consider a student who is asked: "How many batches of spice mix can be made with 9 cups of chili powder if one batch requires ¾ cup?" The correct answer is 12. But a student who writes "27" has made a very different mistake than one who writes "1/12." The first student multiplied instead of dividing. The second inverted the fraction relationship. These are distinct conceptual errors requiring different interventions.

Prior datasets suffered from multiple issues. Some collapsed all of this into a single number: 1 if the student got it right on the first try or 0 if they got the problem wrong. FoundationalASSIST preserves the full picture, enabling a new generation of research that was simply impossible before. Others were limited to multiple-choice problems, meaning real misconceptions could not be teased out, as students had a prescribed set of answers to choose from. Finally, we also include significantly more context per student than any dataset that did include natural text, making learning over time significantly more observable and measurable than before.

Here are four new research directions the dataset unlocks

Problem-level knowledge tracing	Can AI predict whether a specific student will get the next problem right, using the full text of their history rather than just ID codes?
Cognitive student modeling	Can AI go beyond right/wrong and predict the exact answer a student will write? Can AI predict the specific wrong answer they are likely to enter, informing us of a students underlying misconception?

Difficulty Comparison	Can AI judge which of two problems is harder?
Discrimination Comparison	Can AI judge which problems better separate high-vs-low ability students?

What the Initial Results Reveal

We tested four state-of-the-art AI models on the above research questions: GPT-OSS-120B, Llama-3.3-70B, and two variants of Qwen3-Next-80 B.

On the basic task of predicting whether a student will get the next question right, every AI model barely outperformed a trivial strategy of just predicting "correct" every time. The best model achieved 56% accuracy, compared to a baseline of 51%. Further, the models were systematically optimistic: one model correctly predicted success 85% of the time when students answered correctly, but correctly predicted 'incorrect' only 13% of the time. A tutor with this bias would systematically miss the students who need help most.

A separate set of tasks tested whether AI could evaluate the quality of assessment questions. On judging relative difficulty, the models performed reasonably well; the best model correctly identified the harder of the two problems 69% of the time overall, and hit 80% accuracy when the difficulty gap between problems was large. This suggests AI has absorbed some genuine intuition about what makes a math problem hard, likely from years of exposure to educational content online.

Every model performed below random chance when judging which problems best distinguish strong students from struggling ones, suggesting AI has learned the wrong lesson about what makes a problem diagnostic.

Each model failed when measuring a concept called discrimination: the ability to identify which problems best separate students who truly understand a concept from those who do not. A highly discriminating question is one where capable students almost always succeed and struggling students almost always fail; it provides clean, useful information. A low-discriminating question looks similar to everyone, regardless of ability. Every model in the study performed worse than

random guessing on this task, and the gap was not small; one model scored nearly 19 percentage points below chance. We believe the models may be conflating difficulty with discrimination, mistakenly assuming harder problems are automatically more diagnostic. In reality, a very hard problem that nearly everyone fails tells you nothing useful about the difference between students.

These findings reveal where AI needs to improve and establish clear benchmarks for measuring that progress.

Who Should Use This Dataset

Just as the original ASSISTments datasets powered more than a decade of knowledge tracing research, FoundationalASSIST is designed to support the next generation of work: research that asks not whether a student got a question right, but what they were thinking, where they went wrong, and how intelligent LLMs can help.

FoundationalASSIST is designed for researchers and educators working at the intersection of technology and learning. It is available under a CC-BY-NC-4.0 license for research and educational purposes and is available on Hugging Face (https://huggingface.co/datasets/ASSISTments/FoundationalASSIST).

Worden, E., Heffernan, C., Heffernan, N., & Sonkar, S. (2026). FoundationalASSIST: Dataset for Foundational Knowledge Tracing & Pedagogical Grounding of Large Language Models.