Open Source AI Model for Tutoring (EDU AI)

Letter of Intent - *Applicant Instructions*

K-12 AI Infrastructure Program

Grant at a Glance

Detail	Information
Grant Name	Open Source AI Model for Tutoring (EDU AI)
Investment Amount	USD $7–$8M, 1 award anticipated
Grant Period	36 months
LOI Release Date	May 8, 2026
LOI Due Date	May 25, 2026
RFP Release	Mid June 2026
RFP Due Date	Late July 2026
Award Decision	Late August 2026
Grant Start	October 2026
Grant Management	Both the proposal review and award monitoring will be managed directly by the Gates Foundation. The Digital Promise K-12 AI Infrastructure Program will support communication, participation in the K-12 AI Infrastructure community, and dissemination of public goods.
Eligibility	Participation is open to any organization or institution; models must be tested on U.S. K-12 student data.

Project Description and Goals

The purpose of this grant is to fund the creation of open-source, K-12 education-specific Artificial Intelligence (AI) model(s) and supporting research artifacts to enable AI math tutoring that is as effective as human experts.

While current general-purpose frontier models have made leaps in capabilities in tasks like reasoning and coding, they are still relatively low performing in authentic learning and teaching contexts. In particular, they suffer from a "helpful assistant" bias, meaning they are optimized for generic helpfulness, minimizing user effort, and providing quick answers, which directly conflicts with the "productive struggle" that is required for human learning. Further, AI models today do not sufficiently account for the learning needs of students furthest from education opportunities in math, including students behind grade level, English learners, and students with learning differences. This project aims to: (1) address an unmet need in the AI models available for students, teachers, education technology providers, researchers, and AI model developers in the US, (2) establish a long term commitment to prioritizing education in AI development, and (3) create digital public goods

such as evaluation frameworks, benchmarks, and measurement tools to ensure we understand how AI models can better support learning for all students.

We know there are many new and exciting machine learning approaches that can potentially bring dramatic improvements to AI accuracy with tutoring that have not been deeply explored. This includes, but is by no means limited to, supervised fine tuning, reinforcement learning, new model architectures, harnesses, etc. We believe you will have many more ideas to suggest to achieve this goal.

For purposes of this project, we define "AI math tutoring" as a student having one-to-one interactions with an AI in the context of K12 learning environments in the United States with the goal of improving student motivation, engagement, metacognition, and learning in math. To be effective in this use case, AI models must incorporate instructional resources and pedagogical approaches that align to a teacher's instructional plans and teaching philosophies while supporting independent student learning. Like effective human tutoring, they should incorporate multimodal resources that include student drawings, video, audio, and text.

While this project is focused on tutoring in math, the components of creating a successful tutoring experience can be extended to many other educational interactions. The model should be designed in a way that the achievements are modularized, testable, and interoperable with many education technology applications. The project will be iterative and specific components of the solution will be identified early on, through an inclusive design process with teachers, students, and other key stakeholders, and progress on these components measured throughout the project.

The model developed in this project will implement an array of complex tasks with high fidelity to learning science principles, speed, and safety. Based on data provided as context and through interactions with students, we strongly suspect that the model should be able to assess what students know in math and what they misconceive. It must be able to understand and attend to a student's motivational state. To provide and reference high quality worked examples and problems, it should be able to draw from high quality instructional materials. As it draws on these inferences and resources, it should apply the most appropriate pedagogical approach that will work for a specific student potentially mixing scaffolding, direct instruction, and motivational support (both in the short and long term).

Models must perform these tasks at low cost, with latency similar to human tutor responses, at a high rate of accuracy, and with utmost attention to safety, anti-harm, and bias mitigation. We know now there are likely many more capabilities AI models must possess to replicate the learning effects of tutoring in math than we mention here.

Application Instructions and Timeline

Please provide your response to this Letter of Intent (LOI) by **Monday, May 25, 2026**. Participation in the LOI is not required for submission to the RFP but is strongly encouraged.

Have a question about this funding opportunity? Submit your questions here.

Ready to submit your LOI? Submit it here.

Use of Submissions & Note on AI Disclosure

LOI submissions will be used by the Gates Foundation to evaluate interest in this opportunity and may inform revisions to the RFP requirements and evaluation criteria. Responses will be treated

as confidential and shared only with members of the RFP administration team. Applicants may revise or expand upon their LOI responses in any subsequent full proposal.

AI-assisted screening tools may be used during the initial LOI review process. All LOI and full RFP submissions will also receive human review. Applicants are encouraged to disclose whether AI tools were used in preparing their LOI.

Team Composition (2,000 characters)

List participating organizations and individuals identifying which role(s) each fills. Why is your team uniquely positioned to conduct this work? It is totally fine if you have roles that you will fill in the future. If so, where might you need connections to other organizations?

This project will require a diverse set of technical, applied, and research skills to achieve the desired impact. These skills may be present in a single individual or organization, or may be distributed across organizations. Effective teams will be able to fulfill the following roles:

Project Role	Required Experience
ML / AI Engineers	Demonstrated experience in AI model development and improvement projects. Capacity to build or refine a high-performance production-grade AI model.
K–12 Practitioners (Districts, Charter Management Organizations, School Networks)	Experience implementing and managing teaching and learning at scale. Experience developing and implementing instructional materials, teacher coaching, and tutoring initiatives. Capacity to run iterative co-design and model testing in real classroom contexts.
Learning Scientists / Education Researchers	Research methodologies, measurement and evaluation, data analysis in relation to the use of AI in teaching and learning. Deep expertise in math tutoring pedagogy, formative assessment, and learning progressions. Experience systematically identifying what enables the effectiveness of tutoring. Capacity to evaluate AI model improvements from multimodal datasets.
Ed-Tech Product Partnerships	Existing or demonstrable relationships with at least one major tutoring ed-tech provider. Capacity to integrate and test the model in products in real classroom contexts.

We know that this project will be complex, risky, and iterative. Given the range of options and approaches available in AI model development today and the rapidly evolving body of evidence on tutoring, we don't prescribe a specific technical approach that applicants should employ to do the things we describe. As such, we want to use this Letter of Intent (LOI) both to get the word out about the project, to give teams time to mobilize themselves for the RFP, and to learn from you how to better design the project goals, project structure and Request for Proposal.

Tutoring Goalposts (3,000 characters)

What are the key learning and teaching areas that your model will seek to improve compared to current Frontier models? Please suggest revisions, substitutions, additions to the goalpost examples provided in this document.

We are not yet certain what the most appropriate frameworks, metrics, or methods are for defining success and measuring progress in this project. We recognize that there are multiple valid approaches and are seeking applicant input on how this project should best hold itself accountable for meaningful improvement.

As part of your response, please describe the key limitations in current AI model performance in tutoring context and how you would address it. What approaches, metrics, or evaluation strategies would you prioritize, and why?

Below are illustrative failure modes observed in current frontier models. These are examples, not requirements, and are intended to signal areas of interest:

Barrier	Description
Verbosity	Models talk too much and too long, failing to create space for student thinking and response.
Helpful Assistant / Solver Bias	Models give away answers rather than supporting the student's learning process (productive struggle). This is the core 'Helpful Assistant' bias produced by instruct-tuning.
Misconception vs. Slip Recognition	Models fail to distinguish between conceptual misconceptions (requiring re-teaching) and simple arithmetic slips (requiring acknowledgment), responding identically to both.
Affect Detection	Models cannot reliably identify productive frustration vs. student anger or disengagement, and therefore cannot deliver appropriate encouragement or adjust pacing.
Student Modeling	Models fail to incorporate knowledge of the individual being tutored (prior performance, learning progressions, linguistic background, group characteristics) to personalize interactions, or incorporate knowledge of the individual in ways that are ultimately harmful.
Contextual Drift	In long interactions (~45-min sessions), models forget initial pedagogical rules and revert to generic, unhelpful behavior.

Phases & Milestones (2,500 characters)

How would you phase this work into smaller segments and what would be the key indicators of success for these phases? Please suggest revisions to the example phases provided in this document.

Below is an example of how the project could be sequenced, including potential phases, activities, and decision points. This example is intended to provide a starting point project phasing, not a

required structure. We welcome your perspective on how this sequencing should be refined or completely re-designed.

Phase & Timeline	Key Activities / Milestones
Phase 1 Foundation & Initial Signal (Months 1–9)	Activities: relevant dataset curation (NTO Million Tutor Moves, SCALE benchmark data, TeachLM, new expert-annotated data) and initial model tuning. Sample milestone(s): annotated datasets added, model selected, early practitioner review session of results has been conducted with teachers and tutors to confirm pedagogical utility ⚠ *Project may be stopped/reassessed if the dataset is insufficient, base model selection is incomplete, or there is no demonstrable improvement over baseline.*
Phase 2 Iterative Model Training & Evaluation (Months 10–18)	Activities: continued model training and refinement, ablation studies documenting relative lift of different modeling approaches and ongoing evaluations. Sample milestone(s): improved model performance on goalpost tasks, completion of tutor MVP with classroom use, co-design session with practitioners ⚠ *Project may be stopped/reassessed if performance thresholds are unmet or if architecture and model performance results do not demonstrate gains.*
Phase 3 Partner Integration & Field Testing (Months 19–30)	Potential activities: deploy model via standard AI inference stack (e.g., vLLM/SGLang), integration with an ed-tech partner for product-level evaluation and deployment, documentation of implementation playbooks, risk evaluation, and cost analyses. Sample milestone: large-scale pilot testing with students and practitioners ⚠ *Project may be stopped/reassessed if there are no ed-tech partners or real-world testing involved, or encounter critical issues during student-facing deployment.*
Phase 4 Open Release & Stewardship (Months 31–36)	Potential activities: public release (model weights, training data where permissible, fine-tuning playbooks, operational API), transition to university or research organization steward with community governance, and dissemination through AIMS Collaboratory. Sample milestone(s): model adopted by multiple organizations; benchmark established with other model providers, results disseminated. ⚠ *Project may be stopped/reassessed if there is no steward organization identified or if there is a significant blocker to achieve open release.*

Architectural Approach (1,750 characters)

What technical approaches and methods do you believe will best accomplish the goals of this project?

Given the range of options and approaches available in AI model development today, we don't assume a specific technical approach that a model developer will need to employ to do the things we describe. We would appreciate your perspective and advice on what methods will likely fit this project.

To serve as an open source resources, we anticipate this project may utilize an open-weights model architecture that can be locally replicated and provided as a digital public good, alongside a range of possible approaches to post-training, retrieval, and system design (e.g., LoRA, SFT, DPO/RLHF). We do not assume a specific technical path and are seeking your perspective on what approaches are most appropriate. What methods, tools, architectures, and model development approaches do you believe will be most critical to success? In your response, you may wish to comment on tradeoffs across approaches such as post-training techniques, retrieval-based methods (e.g., RAG), and system-level design choices (e.g., agentic or harness-based architectures), as well as how you would balance performance, cost, and iteration speed. Where relevant, describe how you would design experiments (e.g., ablation studies) to understand the relative impact of different approaches (such as fine-tuning, prompt engineering, or retrieval augmentation).

Data Acquisition Plan (1,500 characters)

How will you acquire the data needed to develop and test your approach? Specify intended data sources and key partners involved, and describe the status of any required agreements or permissions.

Intent to Respond (750 characters)

Confirm your team's interest in and availability to submit a full grant proposal. Are there any major constraints that would affect your ability to proceed?

Existing Resources & Projects

This project will build on several existing efforts to improve math tutoring performance; we encourage applicants to design their response assuming they will have active collaboration and use of the datasets, benchmarks, and toolsets created by teams that include (at a minimum) the following organizations:

National Tutoring Observatory | Website
Rene Kizilcec, Purdue University, Principal Investigator
Justin Reich, Massachusetts Institute of Technology, Co-Principal Investigator

The National Tutoring Observatory is building the first large-scale, open-access multimodal dataset of real-world tutoring interactions, linking millions of educator moves to student outcomes to help researchers and developers understand what makes tutoring effective and scalable.

Example of work: Baseline Performance Pipeline Dataset

AI Math Tutoring Benchmark and Open Dataset Project | Website
Susan Loeb, Stanford University, Principal Investigator

This project is building a novel approach to identifying key learning moments in tutoring and developing and interactive benchmark using simulated students interacting with tutoring models to evaluate their effectiveness to promote student learning.

Math Misconceptions Data Challenge | Website
Bethany Rittle-Johnson, Vanderbilt University, Principal Investigator
Scott Crossley, Vanderbilt University, Co-Principal Investigator
Kelley Durkin, Vanderbilt University, Co-Principal Investigator

This project develops annotation schemes and large labeled datasets to detect misconceptions underlying students' math errors, then challenges teams to build algorithms that can identify those misconceptions automatically in digital learning platforms.

Example of work: MAP - Charting Student Math Misunderstandings

Open-Source Multimodal Math Classroom Dataset | Website
Jing Liu, University of Maryland, Principal Investigator

This project is creating an open-source multimodal dataset from mathematics classrooms, combining transcripts, audio, video, and human annotations of key teaching moves, to support AI model development for education researchers and practitioners.

Example of work: Concept-Guided Chain-of-Thought (CGCoT) pairwise annotation tool for systematic text evaluation using LLMs

Language Co-Pilot Project | Website
Dora Demszky, Stanford University, Principal Investigator

Tutor CoPilot is an AI system that provides real-time, expert-level guidance to K-12 math tutors during live sessions, helping less experienced educators apply proven pedagogical strategies and demonstrating improved student outcomes at just $20 per tutor annually.

Example of work: Scaffolding Middle-School Mathematics Curricula With Large Language Models