Open Source AI Model for Tutoring (EDU AI)

Request for Proposals - *Applicant Instructions*

K-12 AI Infrastructure Program

Grant Opportunity at a Glance

Detail	Information
Investment Name	Open Source AI Model for Tutoring (EDU AI)
Investment Amount	Up to $8,000,000 USD | 1 award anticipated
RFP Release Date	June 1, 2026
Proposal Due Date	July 31, 2026
Estimated Grant Period	30 to 36 months
Estimated Grant Start	November 2026
Eligibility	Open to organizations or institutions meeting the Demonstrated Experience and Minimum Scale requirements. See the Eligibility section below for full details.
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.
Application Link	Submit applications using this Qualtrics Form | PDF version (view only)
Contact/Support	grant-support@ld-insights.com

Introduction & Program Overview

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

This RFP follows a Letter of Intent (LOI) issued May 8, 2026. LOI responses were used to evaluate applicant interest, refine RFP requirements and evaluation criteria, and allow teams time to mobilize for the full proposal. Participation in the LOI was not required for RFP eligibility.

Background & Context

AI Math Tutoring Definition: 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.

Efficacy & Evidence: Current general-purpose frontier AI models suffer from a "Helpful Assistant" bias: 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. Research demonstrates this is not a temporary gap: the DrawEduMath benchmark (Baral et al., 2025) shows models are not improving on education-specific tasks as a byproduct of general capability gains, and they perform worst on exactly the tasks most critical to pedagogy: identifying student errors and supporting students who need the most help. Two major studies demonstrate that improvements on base AI models can close this gap: TeachLM (Perczel, Chow, & Demszky, 2025) showed that fine-tuning on 100,000 hours of authentic tutoring data doubled student talk time and matched human-level conciseness; LearnLM (Google DeepMind) showed a pedagogically fine-tuned model outperformed human tutors on supporting students’ knowledge transfer in a randomized controlled trial. Open-weights models now trail frontier models by only 6-9 months in capability, reinforcement learning methodologies have matured and fine-tuning costs have decreased dramatically, and the National Tutoring Observatory (NTO) (Kizilcec & Reich), Allen Institute for AI / Stanford Scale Tutoring Benchmark framework (Knight, et al.), and TeachLM corpus (Perczel, et al.) provide a robust data foundation, making this a propitious moment for this investment.

Relevant Frameworks: Applicants should be familiar with: Targeted Universalism (designing for the most underserved populations as a means of improving outcomes for all); Evidence-Centered Design (aligning assessment tasks, evidence, and claims); the Allen Institute for AI / Stanford Scale Tutoring Benchmark framework (move taxonomies, evaluation rubrics, and the simulated student methodology); and the ML/AI skills and judgment needed to select and apply appropriate training, fine-tuning, and architectural approaches to meet the project's goals. We anticipate that a successful model will require thoughtfully sequenced experiments combining leading-edge approaches.

Proposal Guidance & Evaluation Criteria

The proposal form is organized into the sections below, each corresponding to one or more of the four evaluation criteria. Total narrative character limit: 23,000 characters across all sections. Charts and tables uploaded separately do not count toward the character limit. The guidance provided for

each section is detailed, but proposals need not address every sub-question separately; concise, integrated responses are welcome. Proposals will be reviewed holistically; the proposal submitted by the most qualified team demonstrating the most promising approach and strongest overall value for dollar will be selected.

- Eligibility & Qualifying Experience → Threshold screen only; does not factor into holistic review
- Overview, Targeted Model Improvements, Technical Approach → Criterion 1: Significance
- Team Background, Key Personnel, Data Acquisition Plan → Criterion 2: Assets and Capabilities
- Project Plan, Measurement & Evaluation, Responsible AI & Safeguards → Criterion 3: Project Workplan
- Confirmation and Dissemination Plan, Targeted Universalism, Global Access & Digital Public Goods → Criterion 4: Release and Dissemination Plan

Eligibility & Qualifying Experience

This section will be used to verify that the team meets the Eligibility requirements before the proposal is forwarded for scoring.

Provide a brief statement (*Limit: 2,000 characters*) describing your most relevant prior work. Include:

- Publication titles with dates confirming work predates May 8, 2026
- Dataset or model release names and links where available
- A description of the scale of prior deployment or evaluation (number of students, grade levels, settings)
- Any institutional agreements or data access arrangements relevant to eligibility

Criterion 1: Significance (WHY)

This criterion asks: why is this project important, why now, and why this team? Strong proposals will cite specific evidence for each claim and connect the proposed approach directly to documented limitations in current models.

Overview

Describe your proposed approach to this work: what you will build, how your approach addresses known limitations in tutoring applications of current frontier models – including any additional limitations beyond those described below that your team has identified – and why this approach is feasible now. Draw on the failure modes and capability gaps described in the Background & Context section. *Limit: 3,500 characters.*

Barrier	Description
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 and acts in contrast to the Socratic process of eliciting student thinking that effective tutors demonstrate.

Misconception vs. Slip Recognition	Models fail to distinguish between conceptual misconceptions (requiring re-teaching) and simple arithmetic slips (requiring acknowledgment), responding identically to both.
Sycophantic Drift	The AI tutor progressively shifts from correcting a student's misconception toward agreeing with it, especially under persistence or emotional pressure.
Verbosity	Models talk too much and too long, failing to create space for student thinking and response.
Memory or Awareness of Student Knowledge	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.
Safety Constraint Erosion	The tutor maintains boundaries in normal tutoring contexts but violates them when the conversation shifts frames, such as through roleplay, hypotheticals, or language switching.

Targeted Model Improvements

Identify the specific limitations of current frontier AI models for math tutoring applications that your project will address. For each limitation: (1) cite the supporting research evidence; (2) describe the datasets or training approaches you have or plan to acquire that would improve model performance in these areas; and (3) explain why your approach and team is positioned to make progress in this area. *Limit: 3,000 characters.*

Technical Approach to AI Model Improvement

Describe the specific architectural approach, hypothesized sequence of experiments, any post-training techniques (e.g., SFT, DPO, RLHF), and model workflows your team will use, including any harness or agentic architectures your team plans to explore or has reason to believe would improve tutoring performance. No particular architectural approach is required; reviewers are looking for well-reasoned methodology using recent innovations that are grounded in your team's prior experience. Include the scale of prior fine-tuning work, compute infrastructure, agent harness engineering, and any preliminary results. Generic methodology descriptions without grounding in prior work will not be competitive. *Limit: 3,000 characters.*

Additional Significance sub-questions reviewers will consider:

- Operationalizing Research: What research-based concepts or insights from learning science, math pedagogy, and formative assessment will be incorporated, and how?
- Targeted Universalism: How does the proposed approach serve the focus population, and how might this approach ultimately benefit broader populations?
- Type(s) of Public Good(s): Why is this specific combination of outputs the right choice for the field at this moment?
- Potential for Adoption: Which technical end-users will adopt this? What is the evidence of demand? How are you assessing the cost of model inference for end users?

Criterion 2: Assets and Capabilities (WHAT AND WHO)

This criterion asks: does the team have what it needs to succeed? Reviewers will look for concrete evidence of existing assets and demonstrated team capacity – not aspirational descriptions of what the team plans to develop.

Team Background

Describe how your team, advisory structures, and practitioner partners bring the technical skills required to build a large-scale AI model that makes significant improvements using state-of-the-art approaches, understands the practical considerations of education stakeholders using AI in real-world school contexts, and has the research base to apply findings from prior studies. The team should also include the perspectives of the students and communities this project serves. *Limit: 1,500 characters.*

Key Personnel

All four required roles must be represented: (1) ML/AI Engineers; (2) K–12 Practitioners; (3) Learning Scientists / Education Researchers; (4) Ed-Tech Product Partnerships. For each key person, briefly describe their role and specific experience relevant to this project. At least one major tutoring ed-tech provider must be identified or conditionally committed at proposal submission for Phase 3 integration testing.

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.

Data Acquisition Plan

Reviewers will assess whether the team has realistic, specific plans for acquiring the data needed to achieve the project's aims. All funded developments are subject to open licensing requirements; please refer to the Defining Public Goods section for additional details. Address these sub-areas:

Training Data Overview: Describe the primary datasets for model training and fine-tuning. For each, indicate source, approximate size, grade levels and subject areas covered, and current access status (secured, in negotiation, or planned). *Limit: 1,500 characters.*

Distinguish between:

- Existing open datasets (e.g., NTO Million Tutor Moves, TeachLM corpus, SCALE benchmark data)
- Proprietary datasets you hold or have access to
- New data you plan to collect, with a realistic collection plan

Data Management Plan: Describe how student data will be protected, de-identified, and anonymized; which datasets can be publicly released and under what conditions; how data sharing agreements with existing dataset owners and ed-tech partners will be structured; and your FERPA/COPPA/state law compliance approach. *Limit: 2,000 characters.*

Criterion 3: Project Workplan (HOW)

The overarching question for this section is: what is the justification and evidence that the workplan will confidently and safely produce the public good(s) described under Criterion 1, building on the assets and capabilities described under Criterion 2?

Project Plan

Provide a project plan showing phases, milestones, timelines, and key decision points. A chart or table is acceptable and does not count toward the 2,000 character limit for this section. Address:

- **Kickoff:** How a rapid start will be achieved once funding is available
- **Phases and Milestones:** Steps and workflows showing how work will proceed, with adequate time for each stage. The phasing table below is an example, not a required structure.
- **Centering:** How the focus population will be centered in co-design and testing – not just named as beneficiaries
- **Contingencies:** For each key risk, describe likelihood, impact, and mitigation
- **Roles:** Clear assignment of responsibilities, especially for data management and practitioner engagement

Example Phasing: 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.

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.*

Measurement and Evaluation

Describe your evaluation approach across four dimensions:

1) **Validity / Reliability / Fairness** – how you will demonstrate that the model improves results in the areas of interest, that it repeats the same results over time, and that it performs appropriately for students from different backgrounds;

2) **Safety & Harm Prevention** – how you will identify, test for, and mitigate potential harms to students, including inappropriate content, bias, emotional harm, and misuse. This is a non-negotiable requirement. Proposals that do not include a rigorous, specific safety evaluation plan will not be considered responsive.
3) **Efficacy** – how you will measure student motivation, engagement, and learning outcomes, including pilot testing with real students;
4) **Cost Studies** – how you will track and report implementation costs to support future replication decisions.

The project will coordinate with the AI Tutoring Benchmark project under development by Allen Institute for AI and the Stanford Scale initiative (scheduled for release late Summer 2026); this benchmark should be included as one of the evaluation measures. Also describe: model performance thresholds at each stage-gate; product-level evaluator structure from ed-tech partners; practitioner co-design session outcomes; evaluation specific to the focus population; and resources allocated to data quality. *Limit: 2,000 characters.*

Responsible AI & Safeguards

Describe your plan for data governance, student data protection, and compliance (FERPA/COPPA/state law). *Limit: 1,500 characters.*

Discuss:

- Technical and organizational safeguards for student data confidentiality
- Harm prevention measures and any responsible AI frameworks or standards your team will apply throughout development and deployment

This model will interact directly with children. A comprehensive safety and bias mitigation plan specific to student-facing deployment is required and will be weighted heavily in review. Proposals should address the following key risks: (1) rapid model obsolescence as open-weights models evolve; (2) benchmark dependency if the designated tutoring benchmark is delayed; (3) student safety and bias mitigation in deployment; (4) steward organization selection for post-grant governance.

Criterion 4: Release and Dissemination Plan (OUTPUTS)

This criterion asks: will the outputs be genuinely open, usable, and adopted? Reviewers will look for a credible, specific dissemination strategy beyond stating intent to publish.

Confirmation and Dissemination Plan

Describe how you will release and disseminate all funded developments. Address your open-source release plan (model weights, training data where permissible, fine-tuning playbooks, and operational API); your steward organization approach including selection criteria, institutional commitment to open-source governance, and technical infrastructure for model hosting, and alignment with education research priorities. Also describe how the team will work with the program team on quality assurance prior to release; confirm adherence to required licenses and describe any restrictions

from data usage agreements or ethics approvals; identify documentation (code, demos, guides, API documentation) that will accompany the release; and outline your dissemination strategy, including any data competitions, webinars, conference engagement, or specific partnerships with hyperscalers or ed-tech standards organizations that could drive broad adoption. *Limit: 1,500 characters.*

Aspects of a compelling dissemination plan may include:

- Possible Data Competitions. To engage technical users across many organizations, the program intends to host data competitions around public goods. It would be valuable for teams to produce one or more designs for a data competition as the proposed project moves into its dissemination phase.
- Broad Promotion. The team can describe how they can directly promote the public good to an audience, such as through a webinar, conference presentation, targeted emails or posts in relevant forums, etc.
- Specific Relationships. The team may have relationships to a hyperscaler, education technology standards organization, or other partner whose incorporation of the public good may lead to large scale impacts on student success. If so, describe how those relationships could be activated

Targeted Universalism

Identify specific priority student populations and describe how their needs are intentionally built into model design, training data curation, practitioner co-design, and evaluation from the outset rather than treated as a post-hoc equity consideration. Explain how the proposed public good(s) will support success for every student, and why this work is likely to create meaningful impact for populations that are often underserved. Please be as specific as possible about the student populations and characteristics your team will intentionally prioritize and design for. *Limit: 1,500 characters.*

Global Access & Digital Public Goods

The Gates Foundation requires that all projects ensure Global Access: (a) knowledge gained must be promptly and broadly disseminated; (b) funded developments must be available at an affordable price in support of the U.S. educational system. All funded developments for this project must be released under a license at least as permissive as CC-BY-4.0 (content) or Apache 2.0 (code/models).

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: the National Tutoring Observatory (Kizilcec & Reich), the AI Math Tutoring Benchmark and Open Dataset Project (Loeb), the Math Misconceptions Data Challenge (Rittle-Johnson et al.), the Open-Source Multimodal Math Classroom Dataset (Liu), and the

Language Co-Pilot Project (Demszky). Full descriptions of each are provided in the References section below.

Defining Public Goods

We are investing in public goods that can be adopted by a technical user (i.e., ed-tech developers, AI researchers, school districts, curriculum teams, and AI model developers). The public goods must be modular, foundational building blocks that can be integrated into a variety of tutoring applications and extended by others beyond this project's timeframe. The model should be designed so that achievements are modularized, testable, and interoperable with many education technology applications.

Licensing Terms. The program team has selected Creative Commons (by Attribution) as the recommended license for datasets and knowledge products (e.g., technical reports, research results, technical reports) and Apache v2.0 for software and code (e.g. evaluations, models, applications) funded through this RFP. Other licenses may be negotiated during the pre-award phase, but a persistent requirement is that the resources are available for commercial or non-commercial use.

Types of Public Goods: Proposers should consider the funding levels when deciding what kind(s) of public good to focus on:

Reference Implementation & Open Model Weights	A public-facing reference application that lets users interact with the model in a realistic tutoring context, alongside programmatic API access for developers and researchers. In addition, other released artifacts should include model weights, training and fine-tuning code, data preprocessing scripts, configuration files, model cards, and inference documentation — everything required to replicate, audit, or extend the model. Hosted via an open repository (e.g., GitHub + Hugging Face) under a permissive license, with versioned releases tied to each project phase.
AI Data Pipeline and Testing Harness	A versioned, reproducible repository covering data ingestion, cleaning, annotation, and train/test splits, plus a testing harness with automated runs against each goalpost barrier. Other teams can point the harness at their own models to generate comparable results. The pipeline will build on and contribute back to existing projects developing resources to improve model performance.
Evals and other Measurement tools	A public evaluation suite covering task-level accuracy on the barriers identified and core tutor capabilities (e.g., math correctness, misconception detection) and pedagogical quality (e.g., productive struggle, scaffolding fidelity, affect-appropriate response, session coherence over long interactions), with documented metrics, scoring rubrics, sample items, and reported inter-rater agreement. Evaluation includes subgroup analyses for the populations the program identifies as furthest from opportunity — students behind grade level, English learners, and

	students with learning differences — to surface differential model behavior before deployment.
Research Results & Lessons Learned	Peer-reviewed papers and preprints on (a) which training and architectural choices most improve tutoring effectiveness, (b) differential performance across student subgroups and what mitigates it, and (c) the validity and reliability of the evaluation framework itself. Datasets, code, model cards, annotation protocols, and negative results released under permissive open licenses, with dissemination through the AIMS Collaboratory and both AI/ML and education research venues.

Grant Opportunity & Eligibility

Eligibility

This opportunity is open to any organization or institution. Proposals must have a robust plan to train and test models using data from K12 contexts in the United States and sufficient experience in those contexts to develop a viable tutoring model. Applications should have a documented plan to gather feedback from teachers, school administrators, and other stakeholders.

- **Demonstrated Experience with U.S. Education AI Models:** Lead organizations must demonstrate prior work with large language models and the ability to improve and apply them in the United States. Accordingly, eligible organizations must have, prior to May 8, 2026, at least one peer-reviewed publication and a demonstrated track record of contributing significant digital public goods (e.g., publicly released dataset, open-source model, evaluation artifacts, or comparable infrastructure resources).
- **Minimum Scale:** Prior work must demonstrate deployment or evaluation on real student or user data at meaningful scale. Proof-of-concept or synthetic-data-only work does not satisfy this requirement. Applicants should briefly describe qualifying work in their proposal; the review team reserves the right to request supporting documentation.
- **Partnerships:** Partnerships across organizations are encouraged and may be required to cover all team roles defined in Assets and Capabilities.
- **Scope Limitation:** This opportunity is intended to support foundational infrastructure and shared ecosystem capabilities. Proposals focused solely on point solutions or standalone end-user applications will not be considered responsive. Applicants should understand and align with our commitment to open source principles.

Budget and Period of Performance:

- **Investment Amount:** Up to $8,000,000 USD
- **Anticipated Awards:** 1 award
- **Estimated Grant Period:** 30 - 36 months
- **Estimated Grant Start:** November 2026

Allowable Costs:

- Staff time (including buy-outs)
- Consulting fees and stipends
- Equipment and computational resources
- Travel (if necessary for project activities)
- Legal and IRB review costs
- Data acquisition, annotation, and anonymization; subcontract costs for ed-tech partner integration activities (Phase 3); compute costs not covered by in-kind contributions

Unallowable Costs:

- Lobbying or political advocacy
- Pre-award costs
- Proprietary product development or commercial product enhancements

When and Where to Submit

Submissions will be accepted until July 31, 2026 at 11:59 PM Anywhere on Earth (AoE) via Qualtrics.

What to Submit

Your submission should include a completed Qualtrics form, relevant supplemental materials, and a separately uploaded budget spreadsheet as outlined below.

Proposal form: Complete all questions in the Qualtrics form. The form is structured to mirror Section A of the Gates Foundation Investment Document, so your responses will carry forward directly into the Investment Document if you receive an award. Each question has an individual character limit; total narrative is 23,000 characters. The guidance provided for each criterion is detailed, but proposals need not address every sub-question separately; concise, integrated responses are welcome.

- Qualtrics Form
- PDF Proposal Form (view only)

Supplemental materials: In addition to the form, please provide the following:

- Prior work or early prototypes, such as initial data samples, model cards, demo items, benchmark results, papers, or reports.
- Resumes or CVs for the Principal Investigator and any other key staff.
- Letters of Commitment for any additional organizations, consultants, or ed-tech partners named in the proposal, confirming they have reviewed their role and are willing to serve. At least one committed or conditionally committed ed-tech partner is required for Phase 3 integration testing. These letters are required for documentation purposes only and will not be reviewed by peer reviewers.
- References and citations (not included in the character count).

Budget spreadsheet: Upload your budget as a .XLS or .XLSX file using the template provided, along with narrative justification. Download a copy of the template and save your version with the file naming convention: PI_LASTNAME_EDUAI_BUDGET.XLS.

- Link to forced copy of Google Sheet
- Budget template Instructions

Note on AI Disclosure

AI-assisted screening tools may be used during the initial RFP review process. All RFP submissions will also receive human review. We anticipate applicants will utilize AI in preparing their proposal materials. In the spirit of responsible use of AI, applicants are required to acknowledge if and how AI was used in drafting their responses. A dedicated field is provided in the proposal form.

References

Baral, S., Lucy, L., Knight, R., Ng, A., Soldaini, L., Heffernan, N., & Lo, K. (2025). *DrawEduMath: Evaluating Vision Language Models with Expert-Annotated Students' Hand-Drawn Math Images.* Proceedings of NAACL 2025 (Outstanding Paper Award). https://arxiv.org/abs/2501.14877. Dataset and leaderboard: https://drawedumath.org/

Demszky, D. (Stanford University). *Language Co-Pilot Project (Tutor CoPilot).* An AI system providing real-time, expert-level guidance to K-12 math tutors during live sessions, with demonstrated student outcome improvements. https://edunlp.stanford.edu/projects/tutor-copilot. Example work: https://edworkingpapers.com/ai24-1028

Google DeepMind. *LearnLM.* A pedagogically fine-tuned model demonstrating improved student outcomes over human tutors in a randomized controlled trial. https://cloud.google.com/solutions/learnlm

Kizilcec, R. (Purdue University) & Reich, J. (MIT). *National Tutoring Observatory.* Building the first large-scale, open-access multimodal dataset of real-world tutoring interactions linking educator moves to student outcomes. https://nationaltutoringobservatory.org. Example dataset: https://github.com/National-Tutoring-Observatory/Baseline_Performance_Pipeline

Liu, J. (University of Maryland). *Open-Source Multimodal Math Classroom Dataset.* Creating an open-source multimodal dataset from mathematics classrooms combining transcripts, audio, video, and human annotations of key teaching moves. https://edsi.umd.edu/projects/benchmark-data-collection. Example tool: https://github.com/mlchrzan/pairadigm

Loeb, S. (Stanford University). *AI Math Tutoring Benchmark and Open Dataset Project.* Developing an interactive benchmark using simulated students to evaluate tutoring model effectiveness. https://tutorsim.org/

Perczel, J., Chow, J., & Demszky, D. (2025). *TeachLM: Post-Training LLMs for Education Using Authentic Learning Data.* https://arxiv.org/abs/2510.05087

Rittle-Johnson, B., Crossley, S., & Durkin, K. (Vanderbilt University). *Math Misconceptions Data Challenge.* Developing annotation schemes and labeled datasets to detect misconceptions underlying students' math errors, with open algorithm challenges. https://lab.vanderbilt.edu/live/2024/06/12/math-misconceptions-data-science-competition/. Example dataset: https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/data

Frequently Asked Questions

Answers to frequently asked questions will be updated during the RFP process: https://k12-ai-infrastructure.org/edu-ai-rfp-faq/.