Teaching & Learning (T&L) Benchmarks and Datasets

Request for Proposals - *Grant Proposal Form*

K-12 AI Infrastructure Program

How to use this form

This form is provided for reference only. Submit your proposal via the Qualtrics form linked below and on the RFP Overview.

Please review the full budget instructions document and submit your budget separately using the budget template.

File naming: `PI_LASTNAME_TL_BENCHMARKS_BUDGET.XLS`

For complete grant instructions please visit https://k12-ai-infrastructure.org/tl-benchmarks-rfp/ .

Grant at a Glance

Detail	Information
Investment Name	Teaching & Learning (T&L) Benchmarks and Datasets
Investment Amount	Up to $5.5M USD | 3 awards anticipated; multiple awards to the same respondent are permitted Up to $2M for Adaptive Learning Experiences & Feedback for Students (1 award), up to $1.5M for Lesson and Instructional Planning for Teachers (1 award), and up to $2M for Teacher Coaching (1 award)
RFP Release Date	Jun 1, 2026
Proposal Due Date	July 31, 2026
Estimated Grant Period	18 to 24 months
Estimated Grant Start	November 2026
Eligibility	Open to organizations or institutions meeting the Demonstrated Experience and Minimum Scale requirements. See Eligibility section in the RFP instructions for full details.
Grant Management	Both the proposal review and award monitoring will be managed directly by the Gates Foundation and Learning Commons teams. The Digital Promise K-12 AI

	Infrastructure Program will support communication, participation in the K-12 AI Infrastructure community, and dissemination of public goods. Learning Commons will also support funded projects as a public asset distribution partner, supporting benchmark hosting and dataset deployment.
Application Link	Submit applications using this Qualtrics Form.
Contact/Support	grant-support@ld-insights.com

Contact Details

Lead Contact Name	
Lead Contact Organization/Affiliation	
Lead Contact Email	

Eligibility & Qualifying Experience

Briefly describe your team's most relevant prior work demonstrating eligibility under the Grant Opportunity & Eligibility requirements. Include the following: your prior experience curating large, expert-annotated open datasets in education and building automated evaluations of LLM outputs (both ideally publicly released before June 1, 2026); your pedagogical expertise in the chosen track (primary in math, with in-house or partnered ELA capacity); and the datasets you will produce or aggregate, including sources, annotation schema, scale, and timeline.

Limit: 2000 characters.

Benchmark Track Selection

Select the benchmark this proposal addresses. Teams wishing to apply for more than one benchmark must submit a separate proposal for each, although you may submit duplicate material in sections as appropriate. Please do not refer to other proposals in your submission; all information needed to evaluate your proposal should be provided within each submission.

- ☐ Adaptive Learning Experiences & Feedback for Students ($1.5M)
- ☐ Lesson and Instructional Planning for Teachers ($1.5M)
- ☐ Teacher Coaching ($2.5M)

Project Description

This is the core narrative section. Total narrative word limit: 19,000 characters across all sections below (excluding AI disclosure).

Overview

Describe the proposed work: what you will build and how will your approach address known limitations in K-12 educational AI infrastructure (both the dataset gap and the evaluation benchmark gap) for the use case in your chosen track. Your response should draw on the failure modes and capability gaps described in the Background & Context section. Your proposal should cite specific evidence for each claim and connect the proposed approach to documented limitations in current models, to theory for effective practices in each domain and to the state of available datasets and benchmarks for improving and evaluating AI performance on these practices.

Track-Specific Representative Tasks

Identify the specific limitations of current AI evaluation frameworks in assessing AI model performance on the use case in your chosen track for K-12 math. Please cite the supporting research evidence, describe existing source materials or annotation resources you have or plan to use to develop expert-annotated ground truth, and explain why your team is positioned to build a rigorous benchmark or dataset where others have not.

> ***Adaptive Learning Experiences & Feedback for Students applicants:*** *describe how your benchmark will measure the model's ability to maintain and update a representation of a student's understanding across multiple interactions and use that representation to select the next instructional move, and how the constructs identified will be operationalized and measured.*
>
> ***Lesson and Instructional Planning for Teachers applicants:*** *describe how your benchmark will evaluate the model's capacity to sequence, prune, and adapt instructional materials for a teacher's actual class while preserving rigor and standards alignment, and how the*

benchmark will produce evidence that is actionable for teachers and ed-tech developers rather than only a single aggregate score.

Teacher Coaching applicants: *describe how your benchmark will jointly evaluate the model's analysis of classroom evidence and the coaching conversation it produces about that lesson, with particular attention to how observation-to-coaching coherence will be measured and how the constructs engage with the multimodal data and annotation challenges specific to classroom tasks.*

Limit: 5000 characters.

Team Background

Detail how your core project team, advisors, and partners provide the technical expertise to create a high-quality gold standard dataset and a rigorous, reproducible benchmark. How do they demonstrate an understanding of practical classroom AI use and knowledge of the existing research base to apply prior study findings? The team should also include the perspectives of the students and communities this project serves.

Limit: 1500 characters.

Key Personnel

List all key personnel. All four required roles must be represented: (1) Pedagogical and educational research expertise sufficient to define what quality looks like in the chosen track and to ground that definition in classroom realities; (2) Dataset construction and expert annotation methodology, including annotation schema design, annotator recruitment and training, inter-rater reliability procedures, and rigorous de-identification, sufficient to produce a defensible gold standard corpus; (3) ML and AI evaluation methodology sufficient to operationalize those quality definitions into a rigorous, reproducible benchmark; and (4) Applied education technology research & development expertise to identify practical tasks and evaluations that would be relevant to education technologies.

Name (Organization)	Role Category *	Expertise

Data Acquisition Plan

Source Materials Overview

Describe the source materials and the tasks they will support across the data corpora, benchmark, and evaluators. For each, indicate source, annotations/labels included, approximate size, grade levels and subject areas covered, and current access status (secured, in negotiation, or planned).

Distinguish between: (1) existing publicly available source materials and benchmarks you intend to draw on or extend; (2) proprietary source materials you hold or have access to through existing agreements; and (3) new data you plan to collect, with a practical collection plan.

Limit: 1500 characters.

Data Management Plan

Describe how student data will be protected, de-identified, and anonymized across different modalities, the consent and assent procedures applied at data collection, how data sharing agreements with existing dataset owners and ed-tech partners will be structured, and your FERPA/COPPA/state law compliance approach.

Limit: 2000 characters.

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 character limit. Address how a rapid start will be achieved once funding is available and the phases and milestones showing how work will proceed with adequate time for each stage.

Limit: 2000 characters.

Measurement and Evaluation

Monitoring, Measurement, and Evaluation Plan

Describe your evaluation approach across the following dimensions: (1) Validity, Reliability, and Fairness (i.e., how you will demonstrate that dataset annotations accurately surface the quality differences you have identified, that the benchmark produces stable results across repeated administrations and across new datasets, and that inter-rater reliability for annotations meets documented thresholds (2) Efficacy (i.e., how you will measure whether improvements in benchmark scores correspond to observable improvements in classroom outcomes, including (where possible) pilot evaluations of products whose models have been tuned against the benchmark) (3) Safety and Harm Prevention (i.e., how you will identify, test for, and mitigate potential harms to students, including inappropriate content, bias, emotional harm, and misuse); (4) Data Contamination (i.e., measures taken towards preserving the integrity of the benchmark derived from the dataset); (5) Cost Studies and Efficiency of Reasoning (i.e., cost-per-evaluation and latency characteristics and comparative results between frontier models and open weight models that have been post-trained or context-optimized using the data corpus.)

Limit: 2000 characters.

Responsible AI & Safeguards

Discuss technical and organizational safeguards for student data confidentiality and harm prevention measures, as well as any responsible AI frameworks or standards your team will apply throughout benchmark development and release. A safety and bias mitigation plan specific to student-facing deployment is required.

Limit: 2000 characters.

Global Access

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).

Confirmation and Dissemination Plan

Describe how you will release and disseminate all funded developments. Address your open-source release plan for all digital public goods developed through this project including the documentation, licensing, hosting and promotion approaches that you will make. A plan for long-term sustainability of the leaderboard following the close of this grant should also be provided.

Limit: 1500 characters.

Targeted Universalism

Priority Focus Populations

Identify the varied learner profiles your dataset and benchmark are designed to cover. Describe how those profiles are built into dataset construct definition, item construction, expert annotation, practitioner co-design, and evaluation from the start.

Limit: 1500 characters.

AI Use Disclosure

We anticipate applicants will utilize AI in preparing their proposal materials. In the spirit of responsible use of AI, we are requesting applicants acknowledge if and how AI was used in drafting applicant responses.

Limit: 500 characters.

END OF PROPOSAL FORM

The form is structured to loosely 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. Additional sections in the Gates Foundation Investment Document will be provided to the grantee at grant execution and are not part of this submission.