Teaching & Learning (T&L) Benchmarks and Datasets

Request for Proposals - *Applicant Instructions*

K-12 AI Infrastructure Program

Grant Opportunity 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	June 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 | PDF version (view only)
Contact/Support	grant-support@ld-insights.com

Introduction & Program Overview

The purpose of this competitive Request for Proposals (RFP) is to fund the design, development, and validation of three distinct and independent, open-source K-12 instructional data corpora, and an Artificial Intelligence (AI) benchmark built for each of the following critical use cases:

1. Adaptive Learning Experiences & Feedback for Students
2. Lesson and Instructional Planning for Teachers
3. Teacher Coaching

Each grant will produce a benchmark as a core public good: a fixed set of pedagogically meaningful tasks that AI models attempt under consistent conditions, scored against rubrics that education subject matter experts have validated. The benchmark will be run against all major AI models as part of this project to create a public leaderboard comparing performance; giving developers, researchers, and the field a shared view of where AI for education stands on these key use cases. Performance on individual tasks within the benchmark should be measured by discrete evaluators, which provide automated tests of model performance.

A benchmark is only as good as the annotated data that is used to create it. We are seeking "Gold Standard" datasets that come from authentic learning contexts, are large (n>1,000), are richly annotated for meaningful constructs, include data about learning behaviors and activities, and include learning outcomes measures where possible.

The primary subject area is K-12 mathematics. Each project must also include an English Language Arts (ELA) extension or pilot component. A portion of the data corpus, annotation framework, or evaluator validation focused on ELA should demonstrate cross-subject generalizability of the resulting datasets and benchmarks.

For each of these use cases, competitive proposals will develop all of the following: richly annotated datasets, automated evaluators and benchmarks. These public goods will be released using open licenses to facilitate widescale adoption and use.

Background & Context

For purposes of this project, we define a "Gold Standard Dataset" as a large corpus of authentic K-12 instructional data (including elements such as high fidelity-classroom audio, teacher instructional plans, clickstream data, handwritten student work, student-tutor interactions, coach-teacher interactions, and other lesson artifacts) annotated against constructs grounded in demonstrated education theory and validated through empirical human expert decisions. These corpora are designed as durable, reusable AI infrastructure, applicable across multiple downstream assets: benchmarks, post-training, context optimization, skill.md development, etc.

We define an "AI benchmark" as a rigorous, open-source measurement instrument, built from data within the Gold Standard Dataset, that automatically evaluates whether a machine learning model

can match human expert quality in K-12 math education on a variety of pedagogically-relevant tasks using automated evaluators. The benchmark is the first packaged utilization of the dataset funded under this RFP and the demonstration of its immediate field value. The primary subject area is mathematics, with required ELA extension or pilot components as described in the Program Overview. We seek to overcome several limitations in the current state of K-12 educational AI infrastructure:

1) **Measurement of Pedagogically Relevant Constructs & Tasks** - most existing education benchmarks do not measure model performance on pedagogically meaningful learning and teaching tasks; instead they evaluate model performance on standardized exam questions.
2) **Lack of High-Quality Annotated Data** - the field lacks large, openly licensed, expert-annotated corpora of authentic K-12 instructional data. This data scarcity limits downstream AI assets for education. Furthermore, existing datasets often insufficiently represent students furthest from opportunity (e.g. multilingual learners, students with disabilities, high-poverty communities) which can introduce bias and limit the validity, fairness, and generalizability of AI systems.
3) **Benchmarks Highlight Model Differences Not Capabilities** - most benchmarks are intended to rank model performance and highlight differences; we seek to identify models that are ready for classroom use; those that can achieve human-level performance on relevant tasks.

Benchmark + Dataset Tracks

The sections below outline the use case(s) for each award track, an example of how that dataset and the benchmark built from it relates to education technologies, an initial set of AI tasks, capabilities, and evidence the data corpus could support, and a benchmark in each award track. The tasks, capabilities, and evidence should be considered as a starting point and not a fixed prescription. We invite expansion and revision to incorporate the expertise of applicants.

Track 1: Adaptive Learning Experiences & Feedback for Students(up to $2M)

Use Case

Customizing learning materials and experiences for students based on their demonstrated abilities is one of the earliest uses of AI in education. Generative AI has the potential to radically expand this area of work and adapt at the individual student level with new considerations, more frequently and efficiently, and in new subjects and materials. These adaptive systems use AI to adjust the difficulty, sequence, and type of practice problems in real time based on a student's responses. They provide instant feedback, guide practice toward mastery, and personalize learning trajectories for each student.

This approach operates inside a tight, multi-turn loop: a student attempts something, the system infers what they understand, and the system selects the next move: a hint, a scaffold, a different problem, a worked example, or an explanation pitched at a specific gap. Ideally these systems can help a student to move into their "Zone of Proximal Development" and create a rich learning experience.

Generic AI models perform poorly here for at least two interlocking reasons. They do not durably represent what a particular student knows across turns, and they default to directly providing answers, resolving frustration but bypassing the productive struggle through which learning consolidates. Further, they do not have broader knowledge about the subject area, reducing the accuracy and specificity of misconceptions diagnoses. A high-quality dataset for this track, and the evaluators and benchmark built from it, must capture and measure both how accurately a model tracks a student's evolving understanding and how appropriately it acts on that understanding, including the case in which the best next move is to not provide an additional problem or solution.

For Example

A middle school student is solving a multi-step algebra problem and writes “3x + 5 = 20, so 3x = 25." Current models, asked to help, will typically point out the calculation error and rewrite the line correctly, or restate the rule for isolating a variable. Either move resolves the student's immediate error but bypasses the question of whether the student understands inverse operations. A high-performing model would recognize that this was a conceptual error commonly made by students and would assign learning materials and experiences that address this specific error and then reassess the item. A benchmark or gold standard dataset for this track should surface this type of difference in response. It should measure whether the model is tracking what this student knows across turns, whether the intervention it selects addresses the actual blocker, and whether the model recognizes the cases where holding back is the right call. It should also report these as separable signals, so the field can see whether a model is good at tracing knowledge, good at selecting the next move, or has uneven performance.

Representative Tasks & Evidence

The table below describes constructs the dataset/benchmark and evaluators should cover, what good performance looks like, and further considerations. These items are a starting point identified through prior research on adaptive feedback and knowledge tracing; applicants are invited to expand, revise, or reframe them to reflect their team's expertise and the data they have access to.

Tasks	Potential Evidence	Further Considerations
Real-time knowledge tracing	Predictive accuracy on next-task performance; persistence of state across turns; appropriate recovery from noisy or inconsistent attempts.	Does the approach include a knowledge graph or other high-level representation of the domain (and misconceptions)? What is the unit of analysis (single turn, episode, multi-session arc) and how does the benchmark or dataset capture states that persist or change (both learning and forgetting) across turns? How is tracing accuracy reported separately from intervention quality, so that a model strong at one and weak at the other is legible to the field?

Just-in-time intervention selection	The selected intervention addresses the actual blocker rather than a generic difficulty; differentiated selection across student profiles producing errors on the same content.	How does the benchmark or dataset capture the conditional structure of intervention quality (the right move depends on what the student knows, what they have just tried, and what they are likely to be one step from understanding)? How does it score items where multiple interventions could be defensible?
Misconception-aware task selection	Misconception classification aligns with knowledge graphs or other validated taxonomies that are interpretable; tasks address the specific misunderstanding rather than re-presenting the original problem.	How does the benchmark or dataset establish or build upon a defensible taxonomy of misconceptions? How does it handle errors that are slips rather than conceptual, and errors that have multiple plausible underlying causes?
Distinguish errors due to disengagement	Ability to identify errors due to disengagement surfacing as carelessness, lack of effort, gaming the system, overconfident rapid response, or other behaviors.	Can the system accurately identify this type of activity and take corrective action to not adjust student knowledge estimates when a student's error does not reflect their ability?
Cognitive demand regulation	Model withholds direct answers when scaffolding is more appropriate; offers worked examples or hints rather than solutions; recognizes unproductive struggle.	How does the benchmark or dataset treat the case where the pedagogically correct action is to withhold support, such as when a student is close to a breakthrough and a worked example would short-circuit their reasoning?
Appropriate feedback selection	Feedback is appropriate to the grade level of the problem and the domain; does not leak the solution and has appropriate justification for materials provided.	How does the benchmark or dataset evaluate feedback as a whole? This includes its content, its calibration to grade level, and its restraint, rather than scoring only whether the feedback is technically correct?
Quality parity with human judgement	AI-generated tasks and activities score within an acceptable margin of expert lessons on a recognized rubric for evaluating the fitness of adaptive materials and experiences, under blind expert review.	What recognized rubric or other standards will be used for the parity comparison, and how will expert review be structured to include knowledge estimations and task / activity selection? Is parity reported as a single composite or as a profile across the rubric's dimensions?

Building on the constructs above, describe how your dataset/benchmark and evaluators 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. Your

proposal should explain how the constructs identified (both those listed above and any the team proposes to add or revise) will be operationalized and measured.

Track 2: Lesson and Instructional Planning for Teachers (up to $1.5M)

Lesson and instructional planning is among the most-used and highest-leverage applications of generative AI in K-12 today: teachers routinely use these tools to create lesson plans, build practice sets, scaffold readings, and adapt materials for the students in front of them. The pedagogical challenge is the practical reasoning a skilled teacher uses to move through a curriculum based on what their class actually needs. Generic models can produce surface-fluent lesson plans, but they struggle to sequence coherently against a learning progression, as identified through a Knowledge Graph or other structured knowledge representation, to identify and target lesson-specific misconceptions. They also frequently do not adapt the modality of a task for English learners or students with learning differences without quietly lowering rigor.

A high-quality dataset and the benchmark and evaluators built from it for this track must capture and evaluate whether the model's planning decisions hold up against the kinds of decisions an expert teacher would make and defend. Applicants are encouraged to anchor the dataset and benchmark built from it to work with widely adopted curriculum such as Illustrative Mathematics 360. The benchmark or dataset must also incorporate other high-quality instructional materials (HQIM) and test generalizability beyond a single curriculum, since one that only measures performance against a single curriculum would have limited market utility. For the required ELA extension or pilot component, applicants should anchor to a designated open-licensed ELA HQIM (such as EL Education).

Further, we encourage applicants to anchor scoring in rubrics designed to evaluate the lesson plans teachers actually produce. The MTSS Center's rubric for high-quality middle school math lesson plans is a strong example: it scores plans against evidence-based and high-leverage instructional practices, such as explicit instruction, scaffolded supports, math discourse, and use of formative data, and was built to give teachers a practical quality check on their own materials. Applicants are welcome to adapt it, combine it with other validated rubrics, or propose alternatives, as long as the scoring reflects judgments that expert teachers would use. For the required ELA extension, applicants should propose an analogous lesson plan rubric anchored in an ELA HQIM context.

For Example

An 8th-grade teacher is planning the next week of instruction on linear equations and has a district license to an HQIM curriculum. She gives the AI tool the standard she is teaching and a copy of last Friday's exit ticket showing ⅓ of her students are still shaky on integer operations. She asks the model to suggest how to sequence the next four days. Current models might return a confidently written plan that hits the standard but minimizes the integer-operations gap, trusting that students will pick it up in context. As a skilled teacher with experience in her district's HQIM, she would make a different call: the gap is broad enough to warrant embedding

integer practice inside the new work rather than separating it, and surfacing the underlying misconception driving the gap. She would first look to the existing HQIM to see if it had supporting materials and activities for this issue before seeking or adapting external resources. A benchmark or gold standard dataset for this track should document instructional artifacts relevant to the instructional plan and measure whether the model's plan actually responds to the class evidence it was given, whether the adaptations it makes preserve the mathematical demand of the original task, and whether its claims about which standard a task addresses hold up against expert teacher judgment. It should also consider coherence and integration of curriculum and an adopted HQIM.

Representative Tasks & Evidence

The table below describes constructs the dataset/benchmark and evaluators should cover, what good performance looks like for each, and further considerations. These items are a starting point identified through research on curriculum design, learning progressions, and standards-aligned instructional materials. Applicants are invited to expand, revise, or reframe them to reflect their team's expertise and the data they have access to.

Tasks	Potential Evidence	Further Considerations
Coherent Sequencing & Pathfinding	Recommended sequence respects validated learning progressions; pacing adapts to evidence about the specific class; backtracking to precursor skills is justified by visible student need.	How does the input to the model represent prior learning: what evidence about the actual class is supplied, in what form, and how does the benchmark or dataset distinguish plans that respond to that evidence from plans that ignore it? How does it score items where multiple defensible sequences could follow from the same evidence? Is there a knowledge graph or other model used to represent curriculum sequencing and progression?
Misconception -targeted task selection	Model identifies common misconceptions associated with a standard; selects or generates tasks that surface or remediate those misconceptions rather than a generic practice item.	How does the benchmark or dataset establish a defensible mapping from standards to misconceptions? How does it handle cases where the model selects a task that is plausible but addresses a different misconception than the one the student is actually showing?
Differentiation without losing rigor	The modality of the task changes (visual representation added, language simplified, dual-language support, content adapted for student interest) but the underlying work asked of the student is not oversimplified.	How does the benchmark or dataset evaluate differentiation for different populations such as English learners or students with IEPs in a way that detects when a model has reduced the underlying mathematical demand of a task rather than genuinely adapting it?

Standards alignment & pacing	Alignment matches expert teacher judgment; the model accurately identifies partial matches and flags activities that claim alignment they do not have. Also has awareness of pacing between standards.	How is standards alignment tested against expert teacher judgment at the activity level? How does the benchmark or dataset handle activities that partially align, or that claim alignment they do not have? Are the activities paced in a way that is reasonable compared to human judgement?
Quality parity with human judgement	AI-generated lessons score within an acceptable margin of expert lessons on a recognized rubric for high-quality instructional materials, under blind expert review.	What recognized rubric will be used for the parity comparison, and how will expert review be structured to avoid order, length, and authorship biases? Is parity reported as a single composite or as a profile across the rubric's dimensions?

Building on the constructs above, describe how your datasets and the benchmark and evaluators built from them 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. Your proposal should explain how the constructs identified (both those listed above and any the team proposes to add or revise) engage with how the dataset and resulting benchmark and evaluator will produce evidence that is actionable for the teachers and ed-tech developers who would consume it, rather than generating only a single aggregate score.

Track 3: Teacher Coaching (up to $2M)

Teacher coaching is the most technically demanding of the three tracks: it requires reasoning across modalities (classroom audio, video, transcripts, observation notes), interpreting that evidence against established frameworks for instructional quality, and conducting a coaching conversation grounded in observations. Coaching shares pedagogical DNA with tutoring a student and both depend on diagnostic listening, calibrated feedback, and a deliberate choice about when to suggest and when to ask. Applicants are encouraged to draw on relevant work in tutoring evaluation where it transfers. Effective coaching also requires lesson-specific noticing: an expert coach evaluates instructional moves not in the abstract but against the specific HQIM lesson being taught including its intended student work, discourse structures, and learning objectives. Good coaching ties feedback to how the teacher's practice enabled (or hindered) enactment of that lesson. The distinct challenge for this track is what we call the observation-to-coaching bridge: ensuring that the model's coaching advice is appropriately grounded in the evidence observed in the classroom, and provides the most relevant feedback at that moment, based on observed behavior, rather than offering generic, plausible-sounding feedback that any model could produce without watching the lesson at all.

For Example

A new teacher records a 40-minute lesson from a unit on equivalent fractions, where students are meant to construct number-line representations and compare them to articulate why two fractions can name the same point. They upload it to an AI coaching tool for feedback. During the lesson, the teacher ran an 85% teacher-to-student talk ratio, prematurely intervened when a student started articulating a productive misconception, and skipped the partner-comparison discussion the lesson was built around. Current models typically return generic suggestions like "incorporate more student voice." Instead, a skilled coach identified the bypassed partner discussion as the highest-leverage issue, since it was the core mechanism for students to articulate the target reasoning. They recommended a specific revoicing move ("Say more about why they look different — can you show us on your number line?") to keep that thinking in the room, and framed a concrete goal for the next observation cycle: protecting at least ten minutes for partner discourse in the next lesson. A gold standard dataset and the benchmark built from it for this track should measure whether the model surfaces the discourse pattern, recognizes the specific HQIM lesson and its intended student work, connects observed moves to lesson enactment, and delivers actionable, lesson-grounded coaching rather than general praise.

Representative Tasks & Evidence

The table below describes constructs the dataset/benchmark and evaluators should cover, what good performance looks like for each, and the further considerations. These items are a starting point identified through research on teacher coaching, classroom observation, and multimodal discourse analysis; applicants are invited to expand, revise, or reframe them to reflect their team's expertise and the data they have access to.

Tasks	Potential Evidence	Further Considerations
Multimodal classroom discourse analysis	Evaluation draws on video, audio, and transcripts of the same lesson alongside student work artifacts a coach would typically review (e.g. written work, student-produced representations); accurate diarization and segmentation across these modalities; agreement with established observation protocols (e.g., MQI, CLASS, Danielson) within reasonable bounds.	What classroom evidence will the benchmark or dataset use as input, and what level of multimodal fidelity is required? How will it handle the diarization and segmentation challenges that are known to depress model performance on classroom audio, particularly in classrooms with overlapping speech or non-standard recording conditions? How will it represent and evaluate the integration of student work artifacts (exit tickets, written responses, problem-solving artifacts) with observed discourse?
Actionable feedback generation	Feedback is observational and specific; tied to a small number of high-leverage moves; uses non-judgmental language; an	How does the benchmark or dataset distinguish feedback that an experienced coach would deliver from feedback that is plausible-sounding but generic? How does it handle the trade-off between

	experienced coach would endorse it as something they would actually say.	volume of feedback (many suggestions) and weight of feedback (one or two high-leverage moves a teacher can act on)?
Observation-to-coaching coherence	Coaching utterances cite specific moments or patterns from the observation; advice would change if the evidence changed; an observer can trace each suggestion back to evidence in the lesson.	How does the benchmark or dataset distinguish coaching advice that is genuinely grounded in observed evidence from advice that is plausible-sounding but generic? Does it contain counterfactual lessons where evidence-grounded coaching should change accordingly, so that the benchmark or dataset can detect a model that produces the same feedback regardless of what it observed?
Coaching-stance calibration	The model selects an open question when teacher reflection is more productive; offers a direct suggestion when stakes or time constraints warrant; avoids excessive scaffolding for experienced teachers.	How does the benchmark or dataset evaluate the choice among asking, suggesting, and holding back, since high-quality coaching depends on knowing when to probe versus when to provide an immediate course of action? How are items scored where multiple stances could be defensible?
Longitudinal coaching coherence	Coaching advice builds on prior cycles rather than resetting; the model tracks what the teacher was working on previously and notices progress or regression on prior goals; feedback across multiple observations shows a coherent arc rather than disconnected critiques.	How does the benchmark or dataset represent the multi-cycle nature of coaching, in which an effective coach returns to a teacher with reference to last session's goals? How are items constructed to test whether a model is genuinely tracking the coaching relationship rather than treating each observation as the first? How is "progress" on a teacher's prior goal scored, given that observable change in practice often takes multiple cycles?
Quality parity with expert human plans	AI coaches are scored within an acceptable margin of error on a recognized rubric for teacher coaching, under blind expert review.	What recognized rubric will be used for the parity comparison, and how will expert review be structured to avoid potential expert bias and congruence between expert judgement? Is parity reported as a single composite or as a profile across the rubric's dimensions?

Building on the constructs above, your proposal should describe: how your dataset and the benchmark and evaluators built from it will jointly evaluate the model's analysis of classroom evidence and the coaching conversation the model produces about that lesson, with particular attention to how observation-to-coaching coherence will be measured (including whether feedback is grounded in both observed instructional evidence and the specific HQIM lesson's design and objectives). Your proposal should explain how the constructs identified (both those listed above and

any the team proposes to add or revise) engage with the multimodal data and annotation challenges specific to classroom observation.

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: 19,000 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 your proposal does not need to 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 the investment will be selected.

The table below maps each form section to the evaluation criteria reviewers will apply:

- Eligibility & Qualifying Experience → Threshold screen (not scored)
- Overview, Dataset + Benchmark Track-Specific Tasks & Evidence → 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 is a threshold screen and is not scored by peer reviewers. Reviewers will use it to verify that the team meets the Demonstrated Experience and Minimum Scale requirements before the proposal is forwarded for review.

Provide a brief statement (2,000 characters or fewer) describing your experience and resources in the following areas. Please include:

1. Prior experience producing or curating large, expert-annotated open datasets in education or a closely adjacent domain, ideally publicly released and available prior to June 1, 2026 (RFP release)
2. Prior experience building and deploying automated evaluations of LLM outputs, ideally publicly released and available prior to June 1, 2026 (RFP release)
3. Pedagogical expertise in the use case area of the chosen track, with primary expertise in mathematics and demonstrated capacity (in-house or through partnership) for the required ELA extension
4. Please describe the datasets you will produce or aggregate for the data corpus and use to develop the benchmark, including data sources, annotation schema and scale, and timeline to collect/annotate them (if not already developed)

Criterion 1: Significance (WHY)

Combined Limit (Overview and Track-Specific Prompt): 5,000 characters.

This criterion asks: why is this project important, why now, and why this team? Your proposal should cite specific evidence for each claim and connect the proposed approach directly 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.

Overview

Describe the proposed work: what you will build and how your approach addresses 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. Draw on the failure modes and capability gaps described in the Background & Context section.

> *Domain Validity in K-12 Education & Hallucination Awareness (Track-specific):* Frontier models continue to climb general-purpose benchmarks (MMLU, GSM8K, expert-level reasoning suites) while their performance on authentic educational tasks such as cognitive demand regulation, quality parity with human plans, and multimodal classroom discourse analysis, remains comparatively low and uneven. Datasets and benchmarks anchored in authentic K-12 tasks and grounded in how expert educators define quality and the benchmarks built from them can help close the gap so that progress on benchmarks is reflected in improved student and teacher experiences. These should ensure fundamentally that the knowledge being represented is accurate and is not created through an AI model hallucination. **Please refer to the Benchmark + Dataset Tracks section relevant to your proposal above for more details.**

Criterion 2: Assets and Capabilities (WHAT AND WHO)

Limit: 1,500 characters.

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

Detail how your 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.

Key Personnel

Your proposal should demonstrate four domains of expertise across the team:

(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 built from the corpus.

(4) Applied Education Technology research & development expertise to identify practical tasks and evaluations that would be relevant to education technologies.

These roles may be filled through partnerships or may be cross-functional capabilities in a team. Describe prior experience for each in this section.

> *Track-specific considerations:* subject-matter experts should reflect the track's constructs and data modalities. For Adaptive Learning Experiences & Feedback for Students, expect expertise in intelligent tutoring systems, knowledge tracing, learning analytics, and clickstream/log data at scale. For Lesson and Instructional Planning for Teachers, expect expertise in curriculum design, standards alignment, and rubric-based evaluation of instructional materials. For Teacher Coaching, expect expertise in teacher education, multimodal data processing, classroom observation frameworks, discourse analysis, and multimodal classroom data capture.

Project Role	Required Experience
Pedagogical and Educational Research Expertise	Research methodologies, measurement and evaluation, and data analysis in relation to the use of AI in teaching and learning. Deep expertise in research and evaluation of AI interventions across educational technology platforms. Capacity to define what quality looks like in the chosen track, ground that definition in classroom realities, translate pedagogical constructs into scorable benchmark items, and evaluate whether model outputs align with expert judgment across multimodal evidence.
Dataset Construction & Expert Annotation Methodology	Demonstrated experience designing annotation schemas, training expert annotators, establishing inter-rater reliability, executing modality-specific de-identification (text, audio, video, handwritten work), and managing data acquisition agreements with districts or ed-tech partners. Capacity to

	produce a defensible gold standard corpus that can support multiple downstream AI infrastructure utilizations beyond a single benchmark.
ML and AI Evaluation Methodology	Demonstrated experience in AI evaluation methodology, scoring harness design, and reproducible benchmark construction. Capacity to operationalize pedagogical quality definitions into runnable, validated evaluators, and to design evaluations that produce stable, comparable results across model releases over time.
Applied Education Technology Research & Development	Existing or demonstrable relationships with at least one major ed-tech provider whose product is relevant to the chosen track. Capacity to integrate the dataset, benchmark, and evaluators into product evaluation pipelines and to test model performance in real classroom contexts using product-realistic conditions.

Data Acquisition Plan

Limit: 1,500 characters.

Reviewers will assess whether the team already has access to existing datasets or specific and viable plans to acquire data and develop expert-annotation capacity required to produce the gold standard corpus described in this RFP. The corpus must be designed as durable AI infrastructure with utility beyond the initial benchmark deployment. Applicants must describe how design choices in corpus structure, annotation schema, granularity, metadata, and sampling explicitly support multiple downstream utilizations, including items and held-out evaluation sets for the benchmark, training-quality data for post-training, instructional strategy or methods context for RAG, expert annotations against validated rubrics for LLM-as-judge ground truth and evaluator development, and skill.md or harness construction for agent solutions. The benchmark is the first packaged proof-of-concept utilization; it should not be the only one the corpus can support. Released artifacts may include the source materials (transcripts, classroom recordings, curriculum artifacts, clickstream logs, student work) from which items and training samples are constructed.

All datasets generated or enriched through this project must be licensed for 'permissive re-use' in commercial and non-commercial contexts; as further described below (under "Licensing Terms") the program team has selected Creative Commons (by Attribution) as the recommended license for datasets.

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:

- Existing open datasets (e.g., NTO Million Tutor Moves, TeachLM corpus, TutorSim benchmark data) you intend to draw on or extend
- Proprietary datasets you already hold or have access to

- New data you plan to collect, with a practical collection plan

Applicants should provide a structural description of representative items in the corpus, including data modalities, key data elements (fields), annotation schema, and contextual metadata, at sufficient detail for reviewers to understand what the corpus will entail. This applies whether the dataset already exists or is to be collected. The description may be included inline within the character limit or as a supplemental attachment (see Supplemental Materials).

Data Management Plan

Limit: 2,000 characters.

The datasets funded under this RFP will contain student speech, classroom audio/video, handwritten work, and other artifacts tied to identifiable students and teachers, and public release requires rigorous, modality-specific de-identification. Describe how student data will be protected, de-identified, and anonymized across modalities; the consent and assent procedures applied at data collection; which dataset components can be publicly released and under what conditions, including any tiered access structure (e.g., public, restricted to accredited researchers, held-out for benchmark use) for components that cannot be fully de-identified; how data sharing agreements with existing dataset owners and ed-tech partners will be structured; your FERPA/COPPA/state-law compliance approach; and any third-party review of these procedures.

Criterion 3: Project Workplan (HOW)

Limit: 2,000 characters.

The overarching question for this section is: what is the justification and evidence that the workplan will confidently and safely produce the public goods 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 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 example phasing below is a starting point, 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 for project phasing, not a required structure.

Phase & Timeline	Key Activities / Milestones
Phase 1 Construct Definition and Ground Truth Development (Months 1–4)	Activities: Recruit and train expert annotators, define annotation schema and constructs aligned to quality frameworks, produce the initial gold standard dataset (assuming data already collected), define tasks for the benchmark drawn from the corpus, and create automated evaluators to test for them. For the coaching award track, this phase would focus on the classroom observation component and conclude with a stage-gate review before proceeding. ⚠ *Project may be stopped/reassessed if evaluators are not created, if annotation data accuracy falls below the acceptable threshold, the gold standard dataset fails quality validation, or the classroom observation pilot reveals fundamental misalignment with the coaching framework.*
Phase 2 Multimodal Integration Pilot Validation & Iterative Refinement (Months 4-10)	Activities: Integrate multimodal data sources, develop scoring methodology, further develop and test evaluations that will be used in benchmark and other downstream utilizations of the corpus, annotate additional data. Sample milestone(s): Agreement between expert reviewers or annotators is established across initial constructs, multimodal data sources identified and successfully integrated, scoring methodology developed and documented, agreement validation completed on early constructs while annotation continues in parallel. Evaluations are created and released as public artifacts. ⚠ *Project may be stopped/reassessed if label agreement baselines fall below acceptable thresholds, multimodal data sources cannot be reliably integrated, or scoring methodology fails to produce consistent results across constructs, item discrimination remains below acceptable thresholds after revision cycles, if performance of evaluation benchmarks cannot be brought to acceptable levels, or if SME review surfaces systematic construct misalignment.*

Phase 3 Testing Suite Design, Full-Scale Validation & Baseline Establishment (Months 10-15)	Activities: Public release of the initial dataset corpus, Design the complete testing suite, including evaluators and benchmark, conduct model evaluations across all major frontier and education-specific models, build the leaderboard infrastructure. Sample milestone(s): initial public data release, first public benchmark release and initial leaderboard scores, documentation of downstream utilization pathways (post-training, RAG, etc.) ⚠ *Project may be stopped/reassessed if model evaluations reveal the evaluation testing suite is not usable or inaccurate to existing applications, leaderboard cannot support public release requirements, or initial scores indicate the benchmark is not meaningful across the intended model range, baseline models fail to produce a meaningful performance distribution (floor/ceiling effects), if no ed-tech partner is engaged for real-world testing, or if field deployment surfaces critical issues with scoring validity or construct coverage.*
Phase 4 Open Release & Stewardship (Months 15-18; overlap with Phase 3)	Activities: Public Release, Developer Engagement, and Iteration. Launch the public leaderboard, complete and document the full dataset release, disseminate through AIMS Collaboratory and K12 AI Infrastructure Program, engage ed-tech developers on dataset utilization (post-training, RAG, etc) alongside benchmark adoption, conduct a first iteration round based on field feedback and model performance data. Sample milestone(s): Public leaderboard officially launched with results for all major models, benchmark disseminated through AIMS Collaboratory and K12 AI Infrastructure Program, ed-tech developer engagement initiated with documented participation, first iteration round completed incorporating field feedback and model performance data. Develop plan for long-term leaderboard maintenance (internally or through transition to other team). ⚠ *Project may be stopped/reassessed if developer and community engagement falls significantly below expectations, field feedback reveals fundamental issues with benchmark validity, or iteration data shows model performance results are not actionable for ed-tech improvement, or a long-term plan for leaderboard maintenance is not feasible.*

Measurement and Evaluation

Limit: 2,000 characters.

Describe your evaluation approach for the dataset and the benchmark derived from it. The table below names the core issues your response should engage with.

Issues to Address	Description
Validity, Reliability, and Fairness	Demonstrate that dataset annotations accurately surface the quality differences you care about, that the benchmark produces stable results

	across repeated administrations and across new datasets, and that inter-rater reliability for annotations meets documented thresholds. For the dataset, this means rigorous annotation procedures, validated construct definitions grounded in education theory, and representative sampling across the learning contexts the dataset claims to cover. For the benchmark, this includes controlling for prompt sensitivity, mitigating known biases in LLM-as-a-judge scoring such as position, verbosity, and self-preference bias, and ensuring evaluations surface differential performance across the diverse student populations served by U.S. K-12 schools, with further guidance provided in the Targeted Universalism section.
Efficacy	Describe how you will measure whether improvements in benchmark scores correspond to observable improvements in learning and teaching, including (where possible) pilot evaluations of products whose models have been tuned using the corpus or evaluated against the benchmark.
Safety and Harm Prevention	Describe 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.**
Data Contamination	As models train on increasingly vast web corpora, they inevitably ingest items and answers of popular public benchmarks, which inflates scores and erodes a benchmark's usefulness. Note that public release of the broader dataset is intended and welcome - it supports post-training, RAG/context optimization, evaluator development, and other downstream uses described in the Multi-Purpose Dataset Design row below. The challenge is preserving the integrity of the benchmark derived from the dataset. Your proposal should describe (1) which portions of the dataset are public from release vs. held out for benchmark use; (2) the contamination-detection approach for the benchmark portion; (3) the cadence and methodology for refreshing held-out items as new models are trained on the public corpus; and (4) how applicants will distinguish, in published leaderboard scores, between models that may have been exposed to public corpus content and those that have not.
Cost Studies and Efficiency of Reasoning	A model that achieves high accuracy only through extensive test-time computation may be unsuitable for real-world classroom deployment. Beyond accuracy, applicants are encouraged to report cost-per-evaluation and latency characteristics, so the field can see the trade-offs between quality, cost, and speed for each model evaluated. Where the dataset supports it, applicants are also encouraged to report comparative results between frontier models and open weight models that have been post-trained or context-optimized using the data corpus - this surfaces the dataset's utility for model capability improvement and to support future replication decisions, not just for evaluation.

The project will coordinate with the grantee selected in the Open Source AI Model for Tutoring (EDU AI), which will develop an open weights model customized to improve LLM performance in tutoring

(to be awarded Fall 2026) and 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.

Performance thresholds in this RFP are not fixed targets. The benchmark and evaluators should empirically determine how datasets and labels are evaluated through human review and other validation approaches. In many cases this is conducted by evaluating how much expert humans agree with one another, both at the overall benchmark level and within specific quality sub-constructs, and use those empirical agreement rates as the ceiling against which model performance is measured. However, there are other cases in which alternative metrics to validate the labels may be used.

Your proposal should also describe the uptake metrics by which initial indications of adoption will be tracked over the grant period for both the dataset and resulting benchmark. For example, the number of ed-tech developers using the benchmark or its evaluators in their product pipelines; the number of teams using the dataset for post-training, RAG, or evaluator construction; documented baseline model scores that enable longitudinal tracking as new models are released; and citations and forks of the public data corpus. Also describe: stage-gate criteria, 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.

Responsible AI & Safeguards

Limit: 2,000 characters.

As AI systems take on more autonomous roles in classrooms, the scope of dataset construction and benchmarking expands beyond accuracy to include student safety and the harms that can arise when these systems are deployed at scale. Your proposal should describe how the benchmark addresses these concerns alongside performance, including which risks the benchmark is designed to surface.

Discuss:

- Technical and organizational safeguards for student data confidentiality across dataset construction, public release, and ongoing stewardship
- Harm prevention measures and any responsible AI frameworks or standards your team will apply throughout dataset and benchmark development and release

Your proposal should also address the following key risks: (1) high cost of obtaining annotated data, (2) timing risks from data dependencies and IRB / data-sharing agreement negotiation, (3) dataset and benchmark sustainability and hosting beyond the grant period, (4) re-identification risk in publicly released corpus artifacts; and (5) frontier lab and developer adoption. A safety and bias mitigation plan specific to student-facing deployment is required.

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

Limit: 1,500 characters.

Describe how you will release and disseminate all funded developments. Address your open-source release plan covering: the gold standard dataset (corpus structure, annotation schema, documentation); the benchmark instrument built from the corpus (item bank, scoring harness, baseline results), supporting public goods (automated evaluators leaderboard infrastructure, composable evaluators), and any documentation required for ed-tech developers and model providers to use the artifacts in their own pipelines for benchmarking, post-training, RAG, and skill.md/harness construction; include your plan for long-term stewardship of both the dataset and an annual benchmarking protocol after grant close. Also describe how the team will ensure that all major models are run through the benchmark 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. A plan for long-term sustainability of the leaderboard following the close of this grant should also be provided.

Applicants should plan to coordinate with Learning Commons as a distribution partner. Learning Commons platform supports benchmark hosting and dataset deployment, including data ingestion into its knowledge graph and publication of the data corpus asset in its entirety.

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. We encourage you to include one or more designs for a data competition as the proposed project moves into its dissemination phase.
- Broad Promotion. Describe how you will 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. If your team has relationships with a hyperscaler, education technology standards organization, or other partner whose incorporation of the public good may lead to large-scale impacts on student success, describe how you would activate them.
- Wide-Scale Integration. Describe evidence that the dataset, benchmark, and evaluators are designed to be used by ed-tech developers and AI labs on their own data and pipelines and not only on the gold standard items released with the

benchmark. This ensures product teams can assess quality in deployment-realistic conditions and that frontier-lab teams can use the corpus as a training/tuning resource and the benchmark as a target to hill-climb against.
- Long-term sustainability plan. Provide an approach for the long-term maintenance and sustainability of the benchmark datasets and leaderboard over time.

Targeted Universalism and Learner Profile Coverage

Limit: 1,500 characters.

Identify the varied learner profiles your dataset and benchmark are designed to cover. At minimum, address grade level, prior knowledge, student achievement levels, English learners, and students with learning differences. Describe how those profiles are built into dataset construct definition, item construction, expert annotation, practitioner co-design, and evaluation from the start.

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).

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, you are required to acknowledge if and how AI was used in drafting your responses. A dedicated field is provided in the proposal form.

Resources & Projects to Consider

This project will build on several existing efforts to improve K-12 educational AI infrastructure; 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. Resources are organized by track, but applicants are encouraged to draw across tracks where relevant.

Track 1: Adaptive Learning Experiences & Feedback for Students

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

This project is developing an interactive benchmark using simulated students to evaluate tutoring model effectiveness.

Draw EduMath | Website
Tools for interpreting handwritten student work, diagrams, and spatial reasoning; essential for any adaptive feedback system operating in real classroom conditions.

EDSI dataset | Website
University of Maryland
Naturalistic student discourse, error patterns, and reasoning processes captured in small-group settings at high audio fidelity.

Feedback Prize - Evaluating Student Writing | Website
Georgia State University & Learning Agency Lab
Datasets and code that automatically segments texts and classifies argumentative and rhetorical elements in essays written by 6th-12th grade students. It includes one of the largest datasets of student writing ever released.

Learning Commons/CZI Knowledge Graph | Website
Learning Commons
Openly licensed knowledge graph database of learner competencies, instructional materials, and learning science research outcomes - including a schema of K-12 math learning components with mapped misconceptions; can strengthen misconception-based task selection constructs. *Note: Learning Commons is also a distribution channel for public assets produced via funding - see Confirmation and Dissemination Plan.*

MathFish | Website
Allen Institute for AI, UC Berkeley, Stanford, EdReports
Evaluates LLM alignment to 385 fine-grained K-12 curriculum standards across 9.9K problems; constructs directly relevant for assessing whether AI can identify skills a student is demonstrating or struggling with.

Track 2: Lesson and Instructional Planning for Teachers

Illustrative Mathematics (IM) | Website
Primary anchor curriculum; openly licensed with minimal IP constraints, allowing benchmark developers and evaluators to work with the full corpus without restriction. Applicants should also design for generalizability beyond IM.

MathFish | Website
Allen Institute for AI, UC Berkeley, Stanford, EdReports
Dataset sourced from IM and Fishtank Learning, directly overlapping with this benchmark's design; consult developers on constructs and mechanics to adapt for evaluating skill identification in planning contexts.

Draw EduMath | Website
Student work interpretation (handwritten, diagrammatic, spatial reasoning) applicable to instructional planning contexts.

EDSI dataset | Website
University of Maryland
Classroom observation data providing ground-truth for planning-to-execution alignment.

Track 3: Teacher Coaching

RPPL coaching dataset | Website
100+ district coaching datasets and ~7,500 utterances with a validated annotation framework across five coaching moves dimensions (Object, Perspective, Function, Tone, Stance); human agreement rates of 73–84% and model kappas of 0.63–0.81 on completed dimensions.

EDSI dataset | Website
University of Maryland
High-quality classroom audio with multiple microphone configurations, annotation against established instructional quality frameworks, and linked contextual variables (teacher demographics, pedagogical intent, curriculum positioning, student data); focused on 4th–8th grade math classrooms using HQIM.

National Tutoring Observatory | Website
Cornell University
Complementary data assets to support coaching benchmark development.

Lastinger Center | Website
University of Florida
Complementary data assets to support coaching benchmark development.

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 ed-tech platforms or curricula 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) and Apache v2.0 for software and code (e.g. evaluations, models, applications). Other

licenses may be negotiated during the pre-award phase, but a persistent requirement is that the resources are available for commercial and non-commercial use.

Types of Public Goods: Proposers should consider the funding levels when deciding what kind of public goods to focus on:

Data Corpus	A data corpus/gold standard dataset designed as durable, reusable AI infrastructure with utility across multiple downstream applications - post-training, context optimization, skill.md and agent/harness orchestration, targeted smaller evaluators, and the benchmark itself. While we appreciate the value of synthetic data, we prioritize data from real-world K-12 learning environments in the United States, covering both mathematics and the required ELA extension. Released artifacts may include the source materials (transcripts, classroom recordings, curriculum artifacts, clickstream logs, student work), the annotation schema and manual, the annotated corpus with documented inter-rater reliability statistics, modality-specific de-identification documentation, tiered access pathways where applicable, and a data card describing distribution coverage and intended uses. Data Corpus should have multiple datasets, annotations on multiple measures, and significant variability across data. We will prioritize large, diverse datasets from a variety of learning contexts over smaller datasets with a narrower scope of applicability. Learning Commons will support dataset deployment, including ingestion into its public knowledge graph and publication of the corpus through its developer platform.
Evaluation frameworks	An openly licensed evaluation framework that defines the benchmark measures, the evidence that counts as a valid signal of each construct, and the scoring rubrics that map model outputs to construct-level judgments. Released artifacts may include the construct-definition document, the annotation manual used to produce the gold standard data, inter-rater agreement statistics, item-level metadata, scoring rubrics and the evidence base behind each, and any sub-scoring rubrics. The framework should be modular so that other teams can adopt individual constructs (for example, a misconception-targeting evaluator) without committing to the full framework.
Benchmarks	The benchmark built from the corpus is the first packaged utilization of the dataset funded under this RFP. It comprises the item bank, the scoring harness, and a public leaderboard for top-level scores alongside a private evaluator that ed-tech developers and model providers can run against held-out items without exposing them to contamination. Released artifacts may include the item bank (with clear documentation of which items are public and which are held out), the scoring harness as runnable code, baseline results across a representative range of models, and an automated workflow for periodic refresh of held-out items to mitigate contamination over time. Hosted on an open repository with versioned releases tied to each project phase.

Evaluators	Composable evaluators that operationalize specific aspects of the benchmark and can be embedded in product-level evaluation pipelines by ed-tech developers, model providers, and school districts. Each evaluator is documented with the construct it measures, the rubric it implements, the evidence rules it applies, its agreement with expert judgment, and the conditions under which it has been validated. Released artifacts may include the evaluator code, configuration files, example invocations against the public item bank, documentation of known limitations and out-of-distribution behavior, and integration guidance for the most common product architectures.
Leaderboards	A public-facing leaderboard that tracks model performance against the benchmark over time. The structure should match the best evidence-anchored way of evaluating the constructs in the use case whether through public/held-out item splits, simulated interactive evaluation, blind expert review, or other approaches appropriate to the constructs being measured. Released artifacts may include the leaderboard site itself, the submission workflow and eligibility criteria, the policy for preserving benchmark integrity over time (such as held-out item rotation, contamination mitigation, or other approaches appropriate to the architecture), the cadence at which scores are recomputed, and the governance procedures for adjudicating disputes or corrections to posted results. The leaderboard should be designed for longevity beyond the grant period, with a documented sustainability plan.

Grant Opportunity & Eligibility

Eligibility

This opportunity is open to any organization or institution. Datasets must be sourced from, and benchmarks must be tested on, U.S. K-12 teacher/student data.

- **Partnerships:** Partnerships across organizations are encouraged and may be required to cover all team roles defined in Assets and Capabilities.
- **Demonstrated Experience with U.S. K-12 AI Datasets and Benchmarks:** Lead organizations must demonstrate prior work with large language models and education data and the ability to improve and apply them in the United States. You should understand and align with our commitment to open source principles. Accordingly, eligible organizations must have at least one peer-reviewed publication and a demonstrated track record of contributing significant digital public goods (e.g., publicly released expert annotated datasets, open-source models, 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. Briefly describe qualifying work in your proposal; the review team reserves the right to request supporting documentation.
- **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.

Budget and Period of Performance

- **Funding Range:** Up to $5,500,000 USD across all awards
- **Awards:** 3 awards anticipated; multiple awards to the same respondent are permitted
- **Estimated Duration:** 18 - 24 months
- **Recommended Start Date:** 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

- Meals, snacks, or alcoholic beverages
- Entertainment or lobbying
- 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 19,000 characters. The guidance provided for each criterion is detailed, but your proposal does not need to 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
- Structural description of the potential dataset(s): representative items showing modalities, data elements, annotation schema, and contextual metadata. See Data Acquisition Plan for guidance.
- 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. 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_TL_BENCHMARKS_BUDGET.XLS.

- Link to copy of Google Sheet
- Budget template instructions

References

Akhtar, M., et al. (2026). When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation. arXiv preprint: arXiv:2602.16763.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., & Fernandez-Llorca, D. *AI Benchmarks: Interdisciplinary Issues and Policy Considerations*. In ICML Workshop on Technical AI Governance (TAIG).

Ferrara, E. (2024). *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies*. Sci, 6(1), 3. MDPI.

Glazer, E., et al. (2024). *Frontiermath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI.* arXiv:2411.04872.

Ishida, T., Lodkaew, T., & Yamane, I. (2025). *How Can I Publish My LLM Benchmark Without Giving the True Answers Away?* arXiv:2505.18102.

Lunardi, R., Della Mea, V., Mizzaro, S., & Roitero, K. (2025). *On Robustness and Reliability of Benchmark-Based Evaluation of LLMs*. arXiv:2509.04013.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). *Focus Article: On the Structure of Educational Assessments*. Measurement: Interdisciplinary Research and Perspectives, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

MLCommons. (2025). *AILuminate: AI Safety and Jailbreak Benchmark*. MLCommons. https://mlcommons.org/benchmarks/ailuminate/

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST. https://www.nist.gov/itl/ai-risk-management-framework.

OECD. (2019, revised 2024). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Phan, L., et al. (2025). *Humanity's Last Exam*. arXiv:2501.14249.

Romanou, A., Ibrahim, M., Ross, C., Shaib, C., Oktar, K., Bell, S. J., Ovalle, A., Dodge, J., Bosselut, A., Sinha, K., & Williams, A. (2026). *Brittlebench: Quantifying LLM robustness via prompt sensitivity.* arXiv:2603.13285.

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., & Agirre, E. (2023). *NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark.* In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 10776-10787).

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., & Vosoughi, S. (2025). *Judging the Judges: A Systematic Study of Position Bias in LLM-As-A-Judge.* In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (pp. 292-314).

Stanford HAI. (2025). *The 2025 AI Index Report — Chapter 2: Technical Performance*. Stanford University. https://hai.stanford.edu/ai-index/2025-ai-index-report/technical-performance

UK AI Security Institute. (2024). *AI Safety Institute Approach to Evaluations*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations

Wan, A., et al. (2025). *The 2025 Foundation Model Transparency Index*. Stanford Center for Research on Foundation Models (CRFM).

Yuan, J., Zhang, J., Wen, A., & Hu, X. (2025). *The Science of Evaluating Foundation Models*. arXiv:2502.09670.

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). *ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs.* In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 1950-1976).

Frequently Asked Questions

Answers to frequently asked questions are provided here and will be updated during the RFP process: https://k12-ai-infrastructure.org/tl-benchmarks-faq/.