Proceedings of Machine Learning Research 339:203–210, 2026 IRAISE Workshop at the Festival of Learning 2026

Understanding the Artificial Intelligence Infrastructural Public Goods Needed for K-12 Education: Community Insights from a Request for Information

Keun-woo Lee klee@digitalpromise.org

Jeremy Roschelle jroschelle@digitalpromise.org

Pati Ruiz pruiz@digitalpromise.org

1001 Connecticut Avenue NW, Suite 935 Washington, D.C. 20036

Abstract

Learning at scale increasingly incorporates generative Al assets, with developers of products and services finding it useful to incorporate multiple Al components in their toolchain and to power particular features. Yet, the practices of choosing, tuning, and measuring Al components for inclusion in a broader product remains somewhat ad hoc, particularly with regard to evaluating Al according to learning sciences concepts. This has led researchers, industry, and funders to call for public goods that would be useful as infrastructure: datasets, benchmarks, and models that can measure learning sciences concepts, strengthen their operationalization, and provide guardrails to protect student privacy and safety. At the launch of a funding program related to this call, a public Request for Information (RFI) was launched and received over 100 responses from researchers, industry, and educators all over the world. Here we report on a work-in-progress to analyze the RFI responses. Using the responses, we answer questions that include: What kinds of infrastructural datasets, models and benchmarks are most needed to better include the learning sciences? What dimensions of learner variability should be emphasized? What privacy, safety and data governance issues are most in need of guardrails? This work presents community-informed insights on AI infrastructure in K-12 education as an initial roadmap to transition towards evidence-based public goods that serve all learners at scale.

Keywords: benchmarks, datasets, models, K-12 education, public goods, AI infrastructure

1. Introduction

Today's AI systems are not designed to meet the complex, high-stakes demands of K-12 education (Bastani et al., 2025; Kosmyna et al., 2025). Most tools lack exposure to the real-world data and teaching practices used in classrooms, and most evaluation methods fail to measure what matters for learning for students furthest from opportunity, and as a result often amplifying existing inequities (Holstein and Doroudi, 2021). The field lacks access to high-quality datasets that reflect the pedagogical, cultural, and instructional realities of K-12 classrooms (Celik et al., 2022). Furthermore, AI models used in education are rarely tuned or evaluated on education-specific tasks like scaffolding learning, identifying misconceptions, or orchestrating instruction (Martin et al., 2024). Finally, the education sector lacks common evaluation frameworks for determining when AI systems are equitably effective, transparent, and safe for students (Kucirkova et al., 2026). Without these building blocks, developers and researchers cannot build or refine AI tools that meet the needs of students and teachers today.

Currently, AI integration in education is often driven by general-purpose models that lack educational context, fail to account for learner variability, and offer few guardrails for safety and privacy (Bastani et al., 2025; Kosmyna et al., 2025). In response, a collaborative launched an initiative to fund AI infrastructure public goods—including datasets, benchmarks, and models—designed to support multiple applications of AI in K-12 education. By creating modular open-license public goods, they aimed to lower the barrier for all developers to create learning science-aligned AI products that serve every student. To ensure this infrastructure meets the needs of the field, they issued a public Request for Information (RFI). We obtained access to the RFI responses and analyzed them to answer three research questions (RQ):

- **RQ1:** Which K-12 challenge areas were identified as most primed for transformation through the introduction of AI infrastructure public goods?
- **RQ2:** What datasets, models, and benchmarks do members of the K-12 education community prioritize as essential public goods for building AI tools grounded in the learning sciences?
- **RQ3:** How can existing AI datasets, models, and benchmarks inform the development of K-12 AI infrastructure public goods?

2. Methodology

A public Request for Information (RFI) was disseminated in the form of a Qualtrics survey in November 2025 through blog publications, social media accounts, and email newsletters with messaging inviting edtech developers, researchers, educators, policy leaders, and additional education community members to respond. The RFI was open for 24 days and was designed to solicit input on the essential requirements for K-12 AI infrastructure. The RFI prompt invited comments on four key areas: (1) identifying educational problems that could be addressed via new public goods; (2) determining the types of datasets, models, or benchmarks most needed or "ripe" for rapid advancement; (3) identifying the resources needed to convert existing AI components into public goods; and (4) identifying existing exemplars to inform the work. Respondents provided open-text comments addressing critical gaps in current AI tools, specific technical needs for datasets and models, and recommendations for data governance and student privacy.

The scope of the analysis includes over 100 responses submitted by a diverse global pool of community members, including academic researchers, EdTech developers, and K-12 educators. A distribution of professional affiliations of RFI respondents is depicted in Figure 1. This initial analysis focuses on identifying recurring themes organized around each research question. Researchers leveraged reflexive thematic analysis, a qualitative research method which analyses and interprets patterns in textual data and emphasizes the active role that researchers play in interpreting themes within their unique context (Braun and Clarke, 2022). This methodology instills iterative reflection throughout the process as new interpretations are developed through an inductive process. In this analysis,

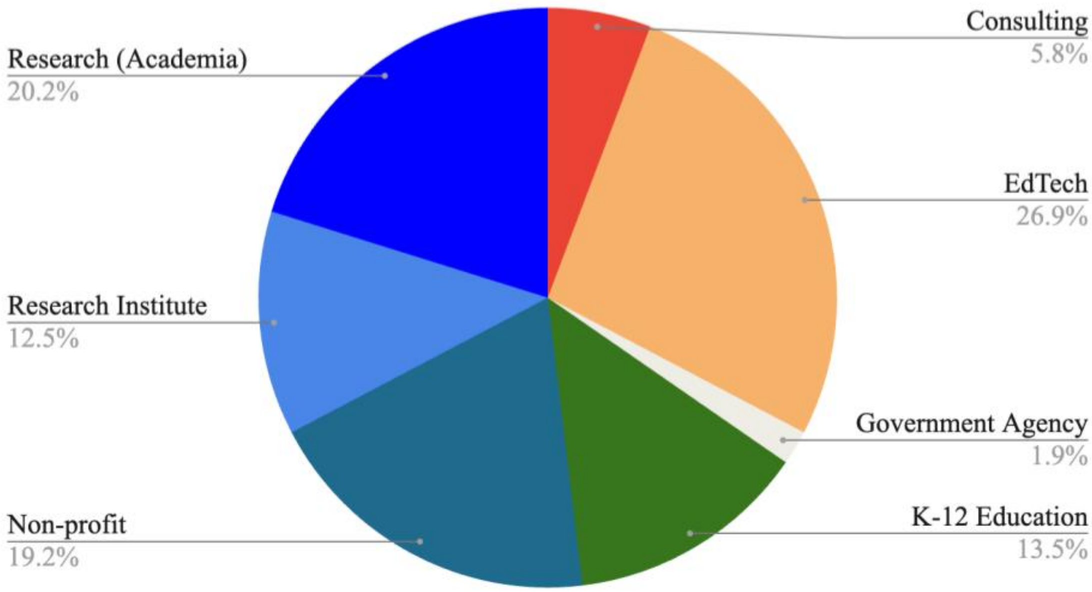

Figure 1: Professional affiliations of RFI respondents ($N = 105$)

researchers included a mix of learning science research expertise, K-12 classroom teaching experience, and computer science knowledge. The Google Gemini Large Language Model was leveraged to augment qualitative analysis, while centering our human expertise and following guidance outlined by Danielle Hitch (Hitch, 2024). To ensure analytical rigor, we employed a multi-stage human-in-the-loop coding process. First, researchers read through the 105 RFI responses to identify an initial set of codes. We then leveraged the Google Gemini Large Language Model to augment the analysis by prompting it to identify common concepts by survey question and across the full dataset. The researchers then performed a comparative analysis, reconciling the AI-generated concepts with the initial human-coded framework and modifying definitions where necessary to ensure accuracy and consistency in terminology usage.

Themes emerged to categorize "Challenge Areas" (e.g., learner variability, pedagogical alignment),"Infrastructural Needs" (e.g., multimodal datasets, children's ASR models), and "Sustainability Needs" (e.g., data governance, open licensing) informed by current barriers with existing AI infrastructure. To ensure the findings represent the needs of all learners, the analysis specifically prioritized insights related to "targeted universalism," focusing on student groups that face the greatest structural barriers, such low-income students, or students 2+ years behind in grade-level (j. powell et al., 2019). These methods ensure that the identified infrastructure requirements are grounded in the practical, pedagogical, and cultural realities of K-12 classrooms.

Table 1: Thematic Mapping of AI Public Goods and K-12 Educational Needs

Pillar	**Core Themes**	**Key Concepts & Action Items**
I. Challenge Areas (The "Why")	Targeted Universalism	Shift from "deficit-framed" gaps to dismantling structural barriers for all.
	Learner Variability	Datasets must reflect the full spectrum (MLLs, diverse dialects, impairments).
	Data Interoperability	Unlock "trapped" data in SIS, LMS, and IEPs to describe unique learning settings.
	Teacher & Student Support	Automate admin tasks; prioritize human relationships; safeguard digital well-being.
II. Infrastructure (The "What")	Multimodal Datasets	Capture progress via speech, artifacts, and context rather than generic internet text.
	Pedagogical Architecture	Build child-specific ASR for noisy rooms; encode "productive struggle" logic.
	Domain-Specific Modeling	Map content to granular skill taxonomies and High-Quality Instructional Materials (HQIM).
	Next-Gen Benchmarking	Evaluate beyond "accuracy"; test for age-appropriateness and predictive validity.
III. Sustainability (The "How")	Legal & Governance	Use PETs (Federated Learning, Differential Privacy) and standardized MOU/Consent templates.
	Technical Capacity	Fund ongoing pedagogical annotation/curation; avoid "one-and-done" project mindsets.
	Sustainable Ecosystems	Use open licenses to prevent vendor lock-in; provide compute credits for non-profits.
	Expert-Led Design	Anchor tools in human judgment; compensate educators for co-design and annotation.

3. Preliminary Findings

3.1. Challenge Areas to Address

The RFI responses highlighted critical gaps where current public goods and datasets fall short and argued for an emphasis on inclusivity and practical utility. Four themes emerged: targeted universalism, learner variability, student data interoperability, and teacher and student supports. We see this fitting together as follows: To truly champion learner variability, new datasets must reflect the full spectrum of students, specifically including multilingual learners and those with language impairments or diverse dialects. This requires a shift toward better student data interoperability, leveraging untapped sources like Student Information Systems (SIS), Learning Management Systems (LMS), and Individualized Education Programs (IEPs) to create more robust descriptions of both the learners and their unique educational settings. Beyond data systems, respondents expressed a need to prioritize the human element of the classroom by enhancing teacher support and digital wellbeing. They acknowledged that AI offers the potential to automate administrative burdens—such as assessment and fragmented data analysis—thereby translating raw information into actionable insights and freeing teachers to focus on instruction and relationship-building. However, as these tools become more integrated, there was a call to ensure safeguards are established to protect student mental health against unmonitored synthetic relationships and emotional distress. Ultimately, there was a theme to move toward targeted universalism; rather than focusing on deficit-framed disparity gaps, AI systems might implement specific strategies to dismantle structural barriers, ensuring the resulting infrastructure allows every student to thrive.

3.2. Infrastructure Needed to Advance K-12 AI

The RFI respondents identified several specialized public goods that the current AI landscape fails to provide, starting with the need for multimodal datasets that accurately reflect diverse student populations and their learning progress. Multimodal datasets may include speaking, listening, drawing, writing, etc., so students can better demonstrate what they know in a variety of formats. Respondents called particular attention to a need for high quality and diverse datasets to develop Automatic Speech Recognition (ASR) models optimized for children's voices in noisy classroom environments.

The survey data elevated that since most AI systems are trained on generic internet text lacking educational context, a core priority involves building foundational data schemas for interoperability. This includes creating machine-readable curriculum contexts and portable learner profiles that allow a student's strengths, accessibility needs, and prior interventions to travel with them safely and privately. RFI respondents called for interoperability to prioritize comprehensive representation to ensure that AI effectively supports multilingual learners and students with disabilities rather than leaving them behind. As we will discuss later, it is nontrivial to represent diverse students while strongly protecting privacy.

Furthermore, there is a significant call for models that integrate learning sciences and effective pedagogy into their core architecture. This includes models that encode learning science principles to guide "productive struggle" and support learning progressions. Respondents also highlighted a need for domain-specific knowledge tracing models that map content and assessments to granular standards and skill taxonomies across different states, ensuring the effective implementation of High-Quality Instructional Materials (HQIM).

Finally, the community emphasized the necessity of benchmarks that move beyond simple accuracy scores to evaluate how well models support learning. This requires rigorous fairness and safety guardrails—such as bias audits to ensure equitable performance across student groups—and age-appropriateness testing to ensure tools can adapt their complexity for a 3rd grader versus a high schooler. Ultimately, these evaluation approaches must verify reliability, check for the reinforcement of misconceptions, and ensure predictive validity regarding actual student learning gains.

3.3. Key Criteria for Public Goods: Privacy, Usability, Quality Assurance

The RFI responses underscore that sustainable educational infrastructure operates as a socio-technical system rather than a mere collection of code and hardware. A foundational requirement identified by the RFI is the establishment of robust governance and legal scaffolding. Given the sensitive nature of student data, critical infrastructure must include privacy-preserving technologies (PETs), such as federated learning, differential privacy, and synthetic data generation, to allow for rigorous research without exposing personally identifiable information. Respondents described that supporting this technical layer requires standardized legal templates, such as MOUs and consent forms, to reduce the administrative burden on district leaders, alongside neutral third-party bodies tasked with auditing model bias and ensuring long-term stewardship.

Beyond security, the community emphasizes that public goods require long-term usability and dedicated technical capacity. Respondents noted these are not "one and done" projects; they necessitate ongoing funding for data curation, de-identification, and the labor-intensive process of pedagogical annotation. To prevent these tools from becoming obsolete, there were recommendations to create open implementation playbooks for resource-constrained districts and dedicated research "testbeds" that allow for real-world A/B testing against actual student outcomes. Findings also suggest that long-term viability is bolstered by shifting incentives, such as providing compute credits for non-profits and utilizing education-friendly open licenses to prevent vendor lock-in and ensure that resources remain accessible and adaptable.

Finally, our RFI response analysis indicates that infrastructure should be anchored by human expertise and rigorous quality assurance. Respondents felt strongly that the effectiveness of public goods relies on human judgment, from expert teacher annotators labeling misconceptions to learning scientists designing priority benchmarks. According to the survey responses, benchmarks should move beyond technical accuracy to measure instructional quality, pedagogical rigor, and predictive validity—specifically evaluating whether an AI's intervention results in student learning gains. By integrating age-aware safety guardrails and compensating educators for their roles in co-designing these tools, the community envisions a transition toward a collaborative governance model. Such a model prioritizes the voices of those historically excluded from the edtech ecosystem, ensuring that AI-enabled infrastructure is safe, transparent, and effective for all learners.

4. Limitations

While this analysis provides initial insights into the various needs for K-12 AI infrastructure, several limitations must be considered when interpreting these preliminary findings. First, the data is subject to self-selection bias. As an open Request for Information (RFI), the participant pool consists of individuals and organizations already deeply invested in the intersection of AI and education. These "early adopters" may not fully reflect the perspectives of a broader community.

Furthermore, there is a trade-off between qualitative depth and generalizability. This work-in-progress relies on a thematic synthesis of approximately 100 qualitative responses, which offer insights into specific technical needs including the necessity for specialized ASR for middle schoolers, for example, but do not constitute a statistically representative survey of the entire K-12 ecosystem. Consequently, while the findings are rich in detail, they should be viewed as a roadmap rather than universal consensus.

A further limitation concerns the gap between expressed and latent needs. Stakeholders are often better positioned to articulate problems they currently experience than to anticipate the infrastructure solutions that are technically possible or pedagogically optimal to best meet their needs. RFI respondents may have disproportionately named incremental improvements to familiar infrastructure such as better ASR models, cleaner datasets, and more interoperable systems, while underspecifying transformative needs they are not yet familiar with.

5. Discussion & Future Directions

While the primary goal of this RFI request focused on technical investments in datasets, models, and benchmarks, responses surfaced critical human-layer requirements that are inextricably linked to the success of any technological deployment. Respondents frequently highlighted that even the most robust, learning-science-informed models will remain underutilized if educators lack the time, training, or trust to integrate them into their classroom workflows. They noted that issues such as AI literacy, K-12 policies, limitations in education research models, and teacher preparedness are vital to scale impact. Although these human-centric factors remain outside the immediate technical scope, they serve as a necessary reminder that infrastructure must eventually be paired with broader professional development frameworks to achieve meaningful impact.

Additionally, since these findings reflect stated rather than revealed needs, future work may include co-design with K-12 educators and students to surface lived pain points and uncover what they most want AI to do for them, grounding infrastructure priorities in practice. These insights can then be triangulated against existing datasets, benchmarks, and models to distinguish gaps addressable by adapting current assets from those requiring new public goods.

There is also a need to continue research and development regarding conditions and technical features that drive measurable improvements in learning. For instance, while multimodal datasets present promising opportunities, the potential of AI systems to enhance multimodal learning analytics (MMLA) through improvement to the technical landscape are yet to be understood. A systematic literature review showed that deeper links to learning theory, pedagogy, and learning impacts remain limited (Mohammadi et al., 2025).

This Work-in-Progress informs the IRAISE community as it transitions from developing isolated AI applications toward contributing to and leveraging a shared, open-source infrastructure. Establishing long-term stewardship remains a priority to ensure that these public goods evolve alongside the rapid shifts in the generative AI landscape, ensuring all students benefit from safe, transparent, and equitably effective learning tools.

Acknowledgments

This paper was developed with support of the K-12 AI Infrastructure Program, an initiative led by Digital Promise.

References

H. Bastani, O. Bastani, A. Sungu, H. Ge, O. Kabakci, and R. Mariman. Generative ai without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025.

V. Braun and V. Clarke. *Thematic Analysis: A Practical Guide*. Sage, Los Angeles, 2022.

I. Celik, M. Dindar, H. Muukkonen, and S. Jarvela. The promises and challenges of artificial intelligence for teachers: a systematic review of research. *TechTrends*, 66(4):616–630, 2022.

D. Hitch. Artificial intelligence augmented qualitative analysis: The way of the future? *Qualitative Health Research*, 34(7):595–606, 2024.

K. Holstein and S. Doroudi. Equity and artificial intelligence in education: Will aied amplify or alleviate inequities in education?, 2021. arXiv:2104.12920.

j. powell, S. Menendian, and W. Ake. Targeted universalism: Policy and practice. Technical report, Haas Institute for a Fair and Inclusive Society, University of California, Berkeley, 2019.

N. Kosmyna et al. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task, 2025. arXiv:2506.08872.

N. Kucirkova, P. H. Uppstad, R. Holeton, D. Lin, A. Clark-Wilson, and A. Chigeda. Evaluating educational technology: Consolidating across multiple impact indicators and rating systems. *Computers and Education*, 240:105467, 2026.

F. Martin, M. Zhuang, and D. Schaefer. Systematic review of research on artificial intelligence in k-12 education (2017-2022). *Computers and Education: Artificial Intelligence*, 6:100195, 2024.

M. Mohammadi, E. Tajik, R. Martinez-Maldonado, S. Sadiq, W. Tomaszewski, and H. Khosravi. Artificial intelligence in multimodal learning analytics: A systematic literature review. *Computers and Education: Artificial Intelligence*, 8:100426, 2025.