UNIVERSITY OF MARYLAND

CENTER FOR EDUCATIONAL DATA SCIENCE & INNOVATION

rppl Research Partnership for Professional Learning

A Framework for Building High-Quality Education Data for R&D in the Age of AI:

The EDSI Dataset and Expert Insights

OCTOBER 2025

Jing Liu | University of Maryland
Brendon Krall | Research Partnership for Professional Learning
Sarah Montana | University of Maryland
Ting-Yu Ariel Chung | University of Maryland
Heather Hill | Harvard Graduate School of Education

Table of Contents

Acknowledgements

AUTHORS

Jing Liu (jliu28@umd.edu) is an Associate Professor in Education Policy and the Director of the Center for Educational Data Science and Innovation at the University of Maryland College Park.

Brendon Krall (brendon_krall@brown.edu) is a Research Project Manager at the Research Partnership for Professional Learning.

Sarah Montana (smontana@umd.edu) is a Doctoral Student in Education Policy at the University of Maryland College Park.

Ting-Yu Ariel Chung (tchung13@umd.edu) is a Doctoral Candidate in Education Policy at the University of Maryland College Park.

Heather C. Hill (heather_hill@gse.harvard.edu) is the Hazen-Nicoli Professor in Teacher Learning and Practice at the Harvard Graduate School of Education.

SPECIAL THANKS

The authors would like to thank the Gates Foundation, the Walton Family Foundation, and the Chan Zuckerberg Initiative for their financial support. The views expressed in this paper are those of the authors and do not necessarily represent the views of the funding organizations. This article benefits from feedback from Wei Ai, Anika Alam, Michael Chrzan, Hannah Rosenstein, Nathaniel Schwartz, and Katie Westone on earlier drafts. We are grateful to Chad Smith of the University of Maryland for graphic design services. We also thank the 22 experts who contributed their ideas through the interviews with our team.

Executive Summary

A PIVOTAL MOMENT FOR AI IN EDUCATION

Almost three years after ChatGPT was released, AI has gone from a buzzword to a reality, now reshaping nearly every aspect of human lives. In education, AI has generated enthusiasm for its potential to provide every student with a high-quality education that meets their individual needs. As new AI models are being developed and released at breathtaking speed, AI-powered edtech tools are also increasingly being incorporated into classrooms. Today, about 40% of teachers have incorporated AI into their regular teaching routines,[1] while 86% of education organizations now use generative AI, the highest adoption rate among all industries.[2]

THE NEED FOR HIGH-QUALITY EDUCATION DATA

How can we make AI a truly transformative, beneficial force in the classroom? While the answer lies in numerous factors, this article focuses on the central ingredient that makes AI so "smart"– data. Generative AI models, such as ChatGPT, Claude, or Gemini, are all trained on vast amounts of internet data. However, internet data are also undifferentiated, lacking the specific context and knowledge required for effective educational applications. In nuanced domains like education, a generic understanding is not enough. For example, it would be extremely difficult for generic AI to understand why a student might have a misconception of fractions and help to support that student's learning. For AI to improve student outcomes, it must have access to education-specific data that reflects the unique dynamics of learning and teaching.

THE CURRENT EFFORT: THE EDSI DATASET

The Gates Foundation, the Walton Family Foundation, and the Chan Zuckerberg Initiative have launched a series of collaborative investments in building large-scale datasets that can support and accelerate data infrastructure for AI R&D efforts in education. In partnership with researchers from Harvard University and Stanford University, the Center for Educational Data Science and Innovation (EDSI) at the University of Maryland is leading an unprecedented effort to build a benchmark classroom dataset between 2025 and 2027. This dataset will be collected and processed in a way that enables model training, benchmarking, tool-building and deeper research into teaching and learning processes.

Our work builds on and extends prior large-scale classroom data collection efforts in the field of educational research. For example, supported by the Gates Foundation, the Measures of Effective Teaching (MET) project collected around 20,000 videotaped lessons from 3,000 teacher volunteers in six urban districts in 2012-2013. During the same period, the Institute of Education Sciences also supported the National Center for Teacher Effectiveness (NCTE) at Harvard University to conduct a three-year data collection effort that captured classroom recordings from approximately 50 schools and 300 classrooms in four districts. These rich datasets, along with other

1 EducationWeek. (2025) More Teachers Say They're Using AI in Their Lessons. Here's How.

2 Microsoft. (2025). 2025 AI in Education: A Microsoft Special Report.

related efforts that might be at a smaller scale, have helped spur extensive research on teachers and teaching and advanced the field of educational research significantly.

With all of these existing datasets, why do we need more education data? And what makes collecting data for R&D in AI and education different from a typical data collection effort for educational or social science research purposes? Just half a year after ChatGPT was released, a convening at Stanford brought together a group of leading researchers, industry professionals, and education practitioners to discuss how language technologies, broadly defined, can be used to support educators. A strong consensus among this diverse group of stakeholders was that collecting high-quality, open-source education data is one of the highest priorities for the field so AI can fulfill its promise in education.[3] Prior datasets, although successful in advancing educational research, lack key features that can meet the needs of R&D in the age of AI. For example, prior datasets often lack high-quality audio that capture student speech clearly, limiting the ability to study how students engage in classroom discourse and their reasoning processes. Such datasets also rarely offer transcripts of lessons–a key data source for AI model training and research that focuses on classroom interactions. Because of the sensitive nature of the personal data in these datasets, the data often cannot be shared broadly and can only be shared through highly secure channels. For example, accessing MET project data requires a complex application and approval process at the University of Michigan.

To contribute to the corpus of large-scale classroom datasets and provide more high quality data for AI R&D, our research team plans to prioritize a few key parameters in our design, including i) high-quality multimodal data that include audio, video, student and teacher survey data, administrative data, and classroom artifacts that are all linked together; ii) naturalistic data that captures the nuances of student-teacher interactions with a specific focus on maximizing student speech quality and student speaker identification, which will allow researchers to connect students' classroom contributions over time and to survey and administrative data; iii) instruction that is rooted in high-quality instructional materials to allow for sufficient observation of high-leverage teaching practices; iv) a data pipeline that fully anonymizes personally identifiable information that enables convenient access to researchers and solution providers. We will also focus on 4th-8th grade mathematics classrooms and set a goal of achieving a sample size of 300 teachers that capture a range of localities and student bodies.

To inform this work and maximize the impact of the dataset we are building, EDSI partnered with researchers from the Research Partnership for Professional Learning (RPPL)[4] to interview 22 experts spanning industry, data science, social sciences, and educational research. In conducting and synthesizing these interviews, we aim to achieve three goals: first, to gather best practices on topics like survey design, privacy, and dissemination, and to ensure our data collection meets the field's needs. Second, to understand the full R&D potential of a dataset like ours by learning how experts might use it, thereby facilitating effective dissemination. Finally, to share the resulting insights with the broader community, collectively advancing the field at the intersection of AI, education, and data science.

3 Demszky et al., 2023, Empowering Educators via Language Technology. Stanford University.

4 https://rpplpartnership.org/

Study Methods

INTERVIEW AND SAMPLING STRATEGY

The EDSI and RPPL research team interviewed 22 experts between March and June of 2025 to guide the development and dissemination of the EDSI dataset. Our interviewees represent a broad range of expertise, disciplinary backgrounds, career stages, and institutions, including technical experts in AI, developers of educational AI tools, and educational and social science researchers who work with large-scale education datasets, AI methodologies, or AI products and services.

We began the recruitment process with purposive sampling by contacting potential participants directly via email. The initial pool was then expanded through snowball sampling, with each expert being asked to provide referrals to other relevant individuals. Two staff members conducted virtual semi-structured interviews with each expert that lasted about 60 minutes, focusing on identifying the R&D activities potential users would pursue, how to best design the dataset's structure and access features, and any related ethical or privacy concerns. All interviews were recorded and transcribed (see Appendix for the full protocol), and participants were compensated $150 for their time.

Once all interviews were completed, the research team conducted a thematic analysis of the interview data using the Dedoose qualitative analysis software. We used a hybrid approach with both inductive and deductive coding, developing a preliminary codebook from the interview protocol to capture suggestions on dataset design, data collection, dissemination strategies, and privacy concerns. We then supplemented analysis with inductive codes that emerged directly from the transcripts.

To ensure reliability, we first trained on one interview from each stakeholder group until inter-coder agreement reached 100%. We then coded the remaining transcripts individually. A second author reviewed each coded excerpt to identify and resolve any discrepancies through discussion until we achieved consensus. Finally, the research team convened to synthesize the coded data, summarizing the major themes that emerged.

INTENDED AUDIENCE

While the EDSI dataset focuses on 4th-8th grade mathematics classrooms, we believe this white paper will benefit a diverse group of stakeholders working at the intersection of AI, education, and data science. We purposely highlight insights in this white paper that are more generalizable across subject domains, teaching and learning contexts, educator and student groups, and academic disciplines. Researchers who are embarking on data collection and/or analysis for AI research and development will find insights for their own research questions. AI developers and edtech providers seeking to design and refine education-specific applications through data-driven approaches will find its contents valuable for designing their own effort for improving product efficacy. Finally, the paper provides critical insights for funders, such as foundations and government agencies, that

aim to make strategic investments in projects designed to strengthen the data infrastructure that underpins the future of AI in education. It is our hope that this article will benefit anyone who intends to invest in, collect, or analyze multimodal education data.

STRUCTURE OF THIS ARTICLE

We start by outlining the potential use cases for a high-quality multimodal classroom dataset across disciplines and industry. Next, we delve into the core of our work, detailing the crucial design considerations and sampling strategies that inform our ongoing data collection process. We then discuss the specific challenges and solutions for capturing high-fidelity audio and video in classroom settings and downstream data processing. Next, we focus on privacy issues commonly faced in education data collection as well as those specific to AI purposes, providing specific discussion on personally identifiable information (PII), the consenting process, as well as anonymization strategies. Lastly, we zero in on dissemination considerations and explore the best approaches to reducing barriers and maximizing the impact of such data. Finally, we provide a summary of our key findings and their broader implications for the future of AI in education.

Potential Use Cases for Multimodal Classroom Data

To help readers appreciate the breadth of possibilities that can be supported by such datasets, we start this white paper by presenting key use cases. We distill them into two broad categories–Teaching and Learning and AI and Data Science, with the first one more concerning the fundamental research questions in the substantive field of education and the latter category that is more technical in nature. These two buckets are not mutually exclusive, as many use cases share substantive and technical aspects. For example, while classroom spatial analysis–examining the physical locations of students, teachers, and materials–is classified as a research question in Teaching and Learning, it directly benefits from the development of classroom-specific computer vision models that are enabled by a dataset like ours.

As Table 1 shows, for Teaching and Learning research, we classify the various topics into student- and teacher-focused research based on the particular subjects on which a topic focuses. For AI and Data Science, improvements in speech technologies such as automatic speech recognition (ASR), conversion of spoken word to text, emerged as a major use case. Benchmarking, a process of evaluating a model against a standardized set of tasks or data, also emerged as a use case of interest. Experts also cautioned against inappropriate uses of such datasets, which we listed specific examples of at the bottom of Table 1. Experts felt strongly that our dataset should reinforce the complex nature of teaching and learning; the knowledge and tools derived from this dataset should only aim to augment, not replace, teachers' work. Experts also emphasized the importance of using the data only for informative rather than evaluative purposes, especially when conducting research about teacher effectiveness.

TEACHING AND LEARNING

Student-Focused Research

- Predict factors of student engagement/success
- Trace longitudinal developmental trajectories for individual student(s)
- Develop novel assessments using naturalistic classroom dialogue that captures student strengths more accurately, holistically, and systematically
- Detect patterns of teacher-student interactions and their association with learning outcomes
- Conduct basic research on student learning manifested in their language use, small group collaboration, and sense making
- Relate classroom characteristics (e.g., on- vs. off-task time) and student learning
- Discern speech and writing cues for diagnosing learning disabilities

Teacher-Focused Research

- Investigate teacher communication and engagement with students
- Document teacher use of technology in classrooms
- Evaluate alignment between a teacher's intended instruction and their actual teaching moves
- Analyze pedagogy and instructional routines and their association with learning outcomes
- Analyze classroom spatial dynamics and various outcomes
- Predict teacher effectiveness with richer data for summative purposes
- Identify effective teacher moves that support students with special needs

AI AND DATA SCIENCE

Speech Technology

- Advance children's ASR using their naturalized speech
- Diagnose children with speech delays
- Identify linguistic biases in ASR models
- Enhance accuracy using video and transcripts to complement audio
- Improve open-source ASR with video/audio/transcripts available
- Improve automatic detection of motivational speech and turnover

Benchmarking and Model Training

- Build, train, and test different models regarding speech (e.g., detecting speech disorder), instruction (e.g., automated feedback on instructional moves), mindset (e.g., scoring algorithm for mindset language), and teacher and student support (e.g., identifying students that need targeted support)
- Serve as benchmarking and evaluation sets for new LLM/machine learning models
- Create synthetic datasets to serve as benchmarks
- Develop synthetic students to test AI tools or for teacher training

INAPPROPRIATE USE CASES AND OR LIMITATIONS

- Build tools to "replace" teachers rather than supplement their expertise
- Use data for evaluative rather than formative purposes

Table 1: Use Cases Discussed by Interviewees

Design and Sampling

Any data collection effort starts with the design of the data structure and sampling strategy. Almost all of our experts emphasized the importance of having a clearly defined target phenomenon of interest and optimizing variation to capture it, but those target phenomena varied widely by field and intended research questions. For social scientists, for example, one expert explicitly said

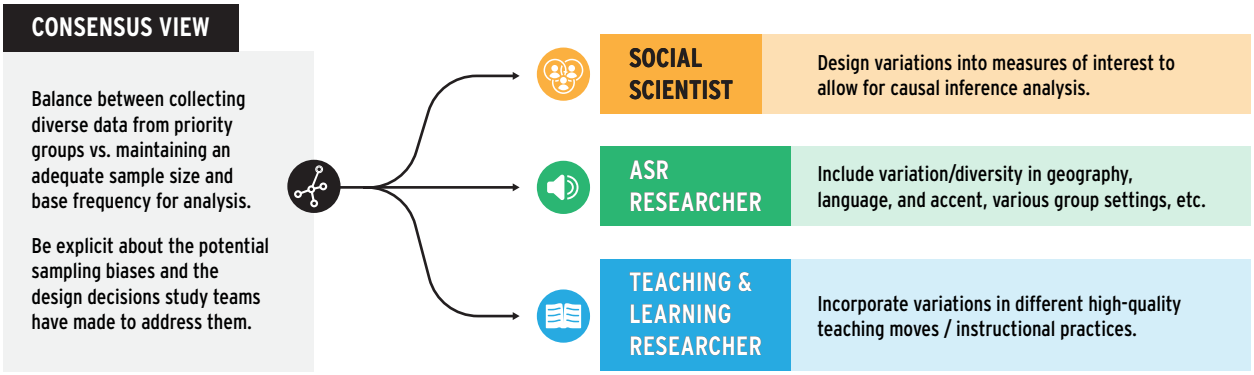

Figure 1: Research Design

to "design variation" into this dataset as naturally-occurring dialogue in classroom settings may not always provide enough variation in the object of study. Another researcher affirmed this need, reflecting on his experience collecting classroom discourse data, saying, "If you just use natural discourse, you're just never going to have certain things that you want to happen with higher frequency. Teachers might not say [the measure of interest] every day." These insights suggest that to study effective teaching, one would need both abundant examples of standards-aligned teaching practices as well as typical practices found in U.S. classrooms sufficiently represented in the dataset. This way, researchers will have the kinds of variation they need to study high-leverage teaching practices that contribute to student learning.

The cost of variation, however, can be statistical power, especially when considering machine learning analysis that relies on big data. For example, one data scientist, in thinking about the dataset's target number of teachers, noted, "With 300 [teachers], it's wonderful, and that's a huge data set. But it's […] actually remarkably small when it comes to some of these [machine learning] methods." Another solution provider also cautioned that "the pursuit of getting a lot of different diverse samples often leads to not having enough samples of any one type […] and that's always a key challenge." Researchers designing data collection for these large multimodal datasets must thus carefully consider these tradeoffs to ensure adequate variation in target phenomena but a large enough sample for meaningful analysis.

In addition to guiding the data collection process, experts across disciplines stressed the importance of clearly defining research interests to help weigh the risks and costs of managing PII in a dataset. Specifically, does one need multimodal data, especially inherently sensitive audio and video, to be able to answer one's research questions? Transforming data post-collection can protect PII but may compromise analytics. For example, student facial data are highly sensitive PII. However, for researchers interested in analyzing student facial features such as eye tracking and facial muscle

movements to measure engagement, high-resolution video is essential, and the common de-identifying practice of blurring faces may not be desirable. But for researchers interested only in eye-tracking data, a technique in which a team documents the coordinate system to map out the main target areas of students' eyesight would allow them to study eye tracking information without keeping data linked to students' actual faces. Making informed choices about data collection and processing for privacy protections requires clarity about specific research purposes.

Having clearly defined phenomena of interest also helps identify potential limitations and sources of sampling bias, or which teachers, students, schools, and districts are represented in the dataset. Experts reiterated the importance of being explicit about the potential sampling biases and the design decisions study teams have made to address them. Being explicit and transparent about limitations and sampling biases is especially important for datasets designed to develop AI-based tools and measures, because products derived from the data will inherently "bake in" those biases. For example, one solution provider reflected on the risk of relying on limited data to develop definitions and measures of good teaching practices:

> *And I don't think a ton of this research has […] taken time to go into different communities and understand […] what ideal classroom culture really looks like and what it means to be responsive [teaching]. With the way that we use these different practices […] we're making a lot of assumptions.*

To help navigate potential limitations and sampling biases in building a dataset, the experts we interviewed pointed out some common aspects of consideration regarding biases. These considerations include diversity of data sources, participants who choose to opt in and out of the study, and the ASR models used to transcribe the data, which can exacerbate biases embedded in the data collection. In thinking about the diversity and what groups might be underrepresented in the dataset, experts across stakeholder groups that work on ASR pointed out linguistic and/or geographic diversity as essential for a multimodal dataset. In terms of the participants opting in and out of the study, one data scientist reminded us to consider "How are the teachers who opted into this unique [study], what is special about them compared to those who opted out?" With regard to bias associated with the ASR model used, our data science experts suggested developing a standardized process to evaluate potential biases introduced by the models. As one data scientist asked, "How do we actually go beyond heuristic-based evaluation and come up with more standardized evaluations of the biases, because there's obviously bias in […] what the model is telling you." Systematic planning and testing like this throughout a project's pipeline, from sampling design to processing, can help transparently document and mitigate bias.

Data Collection

Regardless of the strength of a dataset's design, the quality of the data gathered will determine its usefulness. We heard three themes repeatedly from our experts about investing in data quality. First, gathering rich contextual data enables deeper understanding of phenomena captured in classrooms. In some cases, collecting contextual information can be expensive in time or direct costs, but failing to capture some of those details can compromise the utility of the dataset. Second, investing in technical quality in audio and video data collection is worth the expense, and crisp audio is more valuable than video if a tradeoff is necessary. Lastly, when combining so many data streams into one coherent dataset, it is important to consider synchronization, downstream diarization, and transcription.

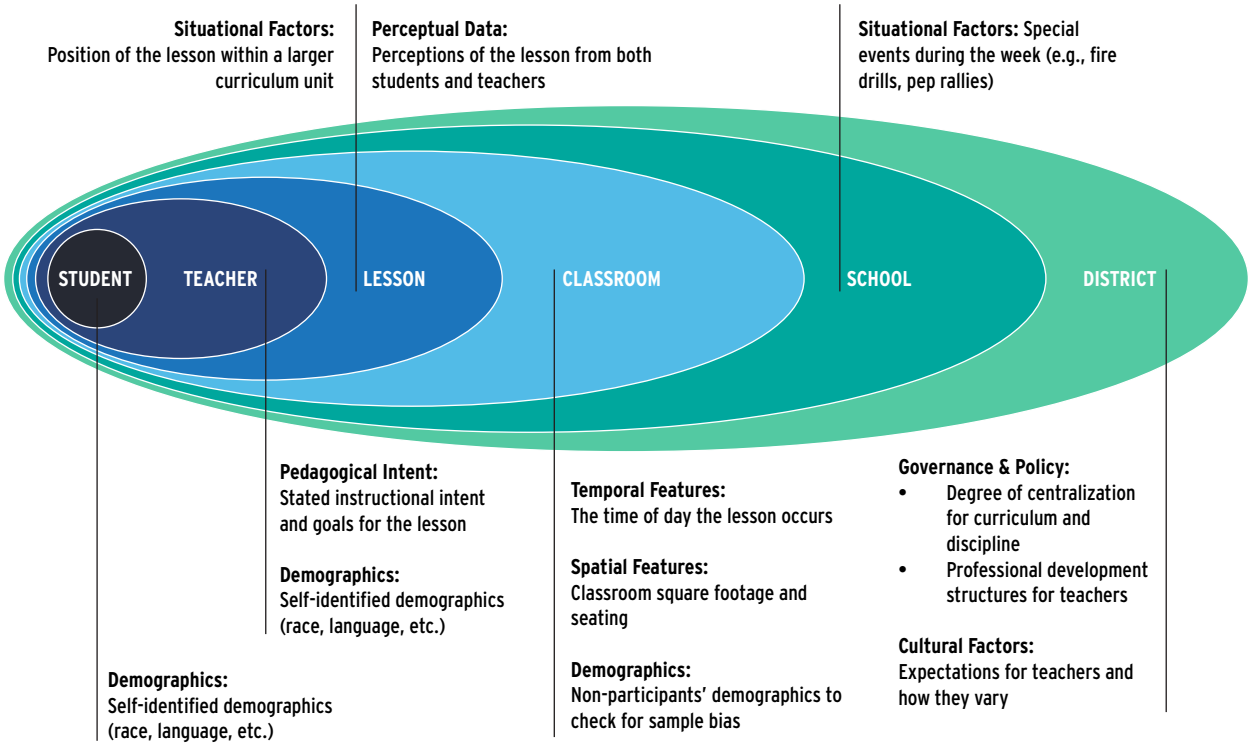

Figure 2: Contextual Variables

CONTEXTUAL VARIABLES

Collecting and documenting contextual factors is integral to data quality.

Experts made clear that a secondary user's ability to accurately make meaning of data is a function of the context available to them. Varied contexts produce varied data, and so understanding sources of variation requires information about students, teachers, curriculum, and school and district contexts. Offering clarity about context extends not only to collecting data, but also explaining the measures used to collect those data. If in doubt, overcommunicate. Our experts offered insights into both what to collect and how to communicate what was collected to ensure that users can confidently interpret classroom phenomena.

Capture rich details of classroom, curriculum, school, and district context, including for those who did not participate.

Teaching and learning researchers offered a broad range of examples of contextual information that can be essential for analyzing classroom lessons, depending on the research questions or potential use cases a dataset may be designed to serve. "Every class is super contextual in many different ways," noted one leading researcher, "Why did the teacher choose to employ certain practices in that particular period? It has many different confounding factors." Collecting and reporting contextual data can help account for those confounders. For example, at the district level, what are the professional learning structures and formal learning experiences teachers encounter? How centralized are curriculum, discipline and other decisions affecting students? How might expectations of teachers differ within and among districts in the sample? At the classroom level, experts suggested gathering information about both spatial and temporal features. For spatial information, measures like the square footage of the classroom and arrangement of student seating can help explain teaching dynamics. For temporal data, information like special events that week (e.g., fire drills, pep rallies, major community events) or simply the time of day of filming can dramatically influence student and teacher behavior. Likewise, capturing lesson-level data like the point in a unit's arc that a lesson falls or the stated intent of a teacher going into the lesson can help explain dynamics as well. Capturing demographic and other data about non-participants is also essential whenever possible, as it provides crucial information about potential bias in the sample.

Whenever possible, offer teachers and students the chance to self-identify personal traits and teachers to share their intention of a lesson.

When gathering demographic information, ask participants to self-identify rather than relying on district or school-provided data, which may be reductive or inaccurate. For example, an Iraqi student may be classified as White according to some district racial categorizations based on outdated Census conventions, or a student who speaks a dialect or a less common language may be only classified as "other language," losing important information. Furthermore, asking teachers and students to share their perceptions of a lesson can help triangulate learning outcomes or other outcomes of interest. Many experts highlighted the importance of understanding teacher intent in lesson activities, especially if their instruction deviates from the written curriculum. What were they intending to do, and why did they make the instructional choices they made? Eliminating guesswork when possible by simply gathering these qualitative data about a lesson strengthens the validity of the data overall.

Document rich details of design and collection of measures, including clear definitions of measures and annotations as well as what is not included in the measure.

While providing rich annotations to go along with the raw dataset will not be feasible for our particular plan (belongs more to data processing than data collection), many experts discussed the importance of high-quality annotations of multimodal data, particularly annotations using common observation protocols (e.g., CLASS, Danielson Framework for Teaching). Aside

from these commonly used rubrics that already contain significant documentation, creating comprehensive and clear documentation explaining the content and processes for collecting key metrics is essential. Clear delineation of the "grain size" of these annotations and measures can also support different user groups. For example, as one learning expert put it, "How do you say what is 'eliciting student thinking'? Literally does it start at minute 2 and 30 seconds in the video? And does it go to 5 minutes, or does the eliciting only [...] into the 3-minute mark, [for example]? Is it the question itself [or does it include student talk]? That's part of what I mean by grain size." Different groups of potential users may also need different levels or types of annotations or documentation of measures. As that same researcher explained:

> *So if AI scientists alone are going to be able to access data, they are going to care much less about the nuances or problems in how the data have been chunked. They are more likely to say 'oh there are these problems with the codes' and just write about that in the limitations section. Whereas learning people [content experts] are gonna have very strong opinions about it because it undermines the validity of the claims that rest on it.*

In this way, clarity in documentation is kindness, and clarity is a path to validity. Providing extensive documentation about what and how data are collected should also include what was not collected. In our previous example, what did not count as eliciting student thinking? If a teacher chooses which of their three course levels to record, which section(s) did they choose not to record? In this case, more information about a measure is generally better. For instance, some experts noted the risks of offering more surface-level data about student engagement without explicitly engaging the complexities of such a holistic concept. "These superficial things like are they moving books around? Do they look like half awake? That's fine. It's like behavioral engagement," they said, "But do you want what they are doing in terms of deep meaning making in their hands and gestures? That entails a certain type of curriculum, […] so I think that's the other important piece here, going back to seeing what am I capturing kids working on [...] and what is the level of analysis I'm interested in?" Documenting the limitations of key measures like those for engagement can help prevent misuse of data, whether intentional or unintentional.

AUDIO AND VIDEO QUALITY

Capturing classroom activity requires careful consideration in terms of microphone and camera placement. The instrumentation process should be purpose-driven data collection with a particular question or use case in mind. At the same time, researchers need to balance what is feasible and scalable with given resources. As one veteran classroom researcher put it, "There's no optimal setup for all classrooms. It really depends on the activity structure of these classrooms, when they implement the curriculum, [and] how much of it involves small group work versus individual work." Given the diversity of classroom setups and potential uses of the data collected in them, our experts offered three guiding principles for collecting high-quality audio and video in classrooms for broad use purposes:

Invest in audio over video quality to support a broader range of use cases when under resource constraints.

Multiple experts questioned the necessity of video data, given the significant logistical and confidentiality challenges involved. The usefulness of video is also highly dependent on the quality of audio data because the latter provides essential context for visual information such as facial expressions or gestures in classroom settings. "Background noise poses a huge problem. Crosstalk poses a huge problem. Differences in mic quality or video quality really matter," noted one expert. The overall takeaway here is that for people who are building datasets under resource tradeoffs, there are many use cases for which high-quality audio can help drive solutions.

Invest in high-quality audio collection for small group discussions for at least a subset of data.

One of the most consistent, and consistently strong, recommendations of our expert sample was to invest in collecting high-quality audio for small group discussions among students. This might include using high-end mic arrays on tables or even mic'ing individual students to get student discourse in small group settings. Without intentional equipment placement, background noise and crosstalk can prevent usable data during student-to-student talk time, which can often be crucial for capturing student learning happening in real time. "I would even do that [mic'ing small groups with high-quality video] well for a subset, rather than badly for everybody," said one leading learning expert, "Being able to identify students and what they're saying to each other is important." Capturing small group audio and video requires careful planning, because placing multiple microphones and video cameras that are not going to get moved can seem infeasible in classroom contexts, and teachers should not be burdened with managing equipment while already managing instruction. Yet the payoffs can be substantial. Investing in high-quality individual microphones for a small subset, as small as 10% of the sample according to one expert recommendation, can also support validation of audio for the larger dataset, allowing calculation of word error rates. To offer one veteran scholar a final word, he argued, "Even if you don't do it at scale, and it's hard to do at scale, you just have to go in there and see what's happening [in small groups]."

Collecting high-quality videos of student facial expressions is both technically and logistically challenging.

Experts agree that video recordings of individual faces are the most logistically challenging and expensive to collect, and data use cases that rely on facial expressions are highly dependent on the quality of such videos. One expert who had experience with video data collection pointed out that even using multiple cameras cannot guarantee close-up data on individual students over the course of a lesson "because these kids move around a lot, and it is very rare to actually get them in the frame," which compromises the possibility of adding annotations or using analytical tools. For nuanced analysis like facial affect coding that examines minute muscle movements in faces, even multiple cameras may not be enough. "Some of the algorithms we apply to multimodal data, when we have machine learning algorithms, they are not working with low resolutions," said another data scientist. Overall, unless facial analysis is a priority use case, many experts recommended

focusing on audio quality at the expense of video given the difficulty of capturing facial features accurately and consistently with enough resolution for analysis.

TECHNICAL CHALLENGES IN PROCESSING DATA

Despite breathtaking advances in processing capabilities over just a few years, technical challenges in preparing classroom data for use are often bigger than one would imagine. We heard four major types of technical problems, all of which can be mitigated to some extent by careful design.

Occlusion

All microphone and camera configurations deal with issues related to establishing "ground truth" (confirmation by direct observation) and occlusion–the loss of data clarity that occurs when participants are blocked from view or sound by movement and objects. In a real-time classroom setting, students and teachers move frequently and background noise or crosstalk between students is common. Often the cost of obtaining such authentic environments is losing quality. "We've realized that pretty much all of these multimodal sensors come with occlusion issues," said one AI expert, "We spent quite a bit of time establishing reliability [...] to have some sort of ground truth to the data." Some degree of omission or occlusion is unavoidable in classroom settings, but careful microphone and camera placement can help ensure a threshold level of usable data.

Synchronization

In order for various streams of temporal data to align, researchers must create a linkable variable for the various data streams, but arriving at the optimal granularity for synchronization can be difficult. Aligning timestamps for audio and video alone can be a technically complex process. Starting with a small amount of pilot data to troubleshoot syncing issues can help head off more major headaches later.

Transcription and Diarization

Despite advancement in ASR systems in recent years, challenges persist in obtaining accurate transcription and diarization of children's speech, especially in noisy classroom settings. "Algorithms are still far from satisfying what we are doing right now, still far from enough in classroom settings and in children's voice," noted one AI specialist, and the nature of multiple speakers in a classroom setting, the "crosstalk" problem, can make that even more challenging. More difficult still is accurately capturing children's voices speaking in accents, especially less common accents. Too often these students' voices are omitted in analyses because algorithms cannot yet discern them, creating potential inequities in the data.

Processing Power

Of course, all of these challenges can also be a function of the processing resources available to

tackle them. Sophisticated models can crunch multiple streams of data, but as one data scientist noted, these models are "pretty expensive, and they are extremely good at convincing you they're right, they hallucinate." As the field progresses, these costs may diminish, but current processing power limitations remain.

Overall, this section discusses the nuts and bolts of what it takes to collect a complex dataset like ours. The rich insights provided by the interviewees about the importance of contextual factors directly informed our own study design, and hopefully will be similarly informative for other similar data collection projects. The technical details in terms of collecting data in noisy and busy classroom settings need to be considered in the context of specific use cases downstream and privacy concerns. Lastly, data processing of multiple streams of data in different modes needs to be planned ahead to mitigate challenges post data collection.

Privacy Considerations

Collecting classroom data inherently carries privacy risks, but the rapidly evolving capabilities of AI-based tools are creating new challenges to contemporary norms for managing personally identifiable information (PII) and consent. Experts we interviewed shared a common concern for privacy and confidentiality, and in particular agreed on the wisdom of offering differing levels of access based on the sensitivity of data, the expanding definition of PII over time, and the

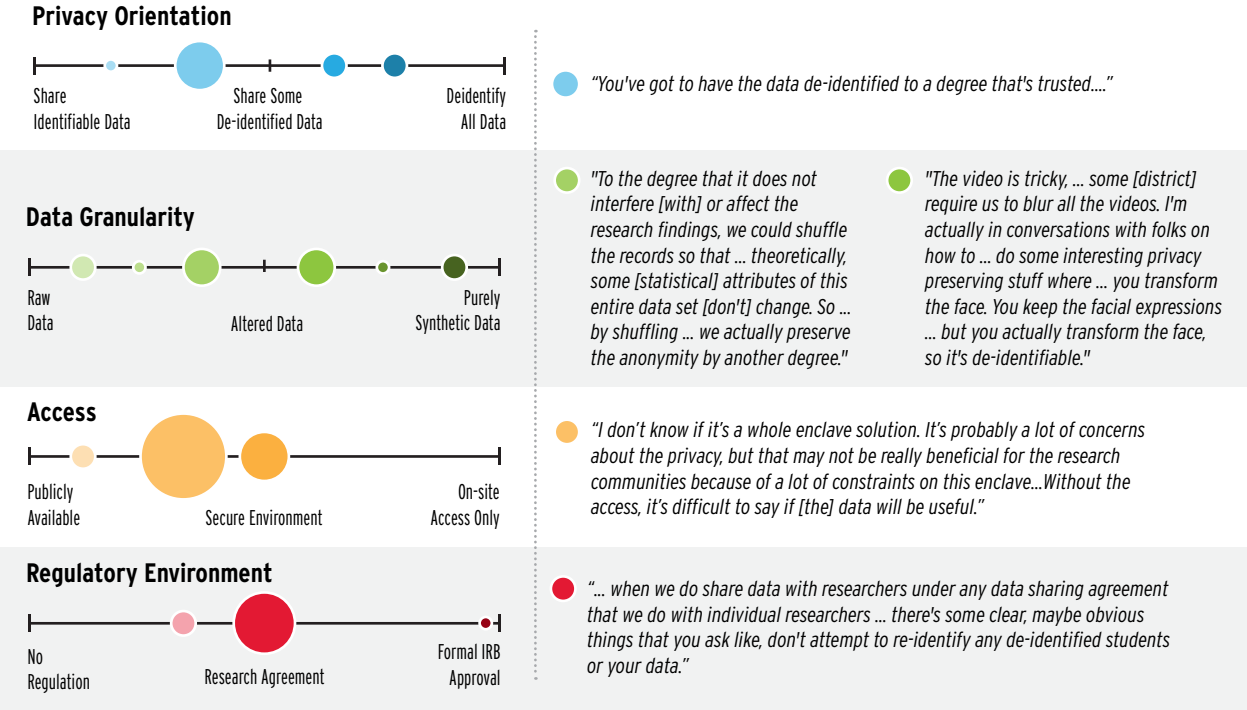

Figure 3: Privacy Orientation. Larger circles indicate more interviewees at that position on the spectrum.

importance of clearly communicated parameters in obtaining consent. However, they differed widely in their degree of preferred privacy protections for both data processing and data access for potential users, and they chose to emphasize differing aspects of security as important. Their spectrum of perspectives is indicative of the emergent nature of these debates.

For example, one AI researcher surfaced the challenge of combatting misinformation related to how private information may be used or misused. "I think one of the biggest concerns is not even privacy, but the perception of privacy," he noted. Offering clear and evidence-based education to participants about what is and is not currently accessible to potentially nefarious actors is important for establishing informed consent. Yet other experts additionally cautioned against underestimating the actual technical risk. One learning expert, advocating for extreme caution, warned that "If someone is really determined even a data enclave is not going to help," and cautioned against releasing particularly audio and video data at all. This stance is driven by the significant re-identification risk posed by ill-intentioned individuals who could identify schools, teachers, or students and use that information to cause harm. Another expert noted that "it's been such a hard thing to make our data publicly available, because it's very hard to give a 100% guarantee that your data is fully anonymized," and counseled against releasing even raw transcripts. The divergence of expert opinion about issues of privacy and security suggests that a conservative approach to privacy offers the most prudent protection against potential violations.

Expect the definition of PII and protection standards to expand over time, and treat potential PII accordingly.

As one data scientist noted, "I think the most worrisome part [of new AI tool uses] comes in around surveillance, let alone security, privacy, and agency." As rapidly improving tools enable the recovery and use of identifiable data, what once was thought safe may no longer be so. Several experts mentioned voice as a prime example of this evolution. "When I was doing this work in 2017, voice was not considered PII, because technology did not exist to remap voice back to an individual," recalled one edtech provider, "And then that technology came in. And now it's very much accepted that voice is PII." Even with modern processing techniques that transform voice, it may be possible with new tools to reverse engineer voiceprint information to identify speakers. Anonymity becomes all the more difficult in a complex project piecing together multiple streams of data on individuals, which increases the chance of triangulating identity. Researchers creating such a dataset might need to take additional precautions to separate and anonymize streams compared to those collecting only voice, for example.

Clearly communicate risks and uses while obtaining consent.

Given the landscape of uncertainty surrounding PII protection amidst expanded AI capabilities, many interviewees discussed the role of a clearly crafted consent form and conversations around it as a tool for responsible management of privacy for participants. A consent form should clearly delineate a timeline (i.e., how many years data can be used and when it will be destroyed), potential uses, and a process for rescinding consent (if possible). Participants deserve clear and

transparent communication about the potential difficulty of deleting data once publicly released as well, if aspects of it are released publicly. As one expert noted, data often cannot be reclaimed once it is released and shared across different universities or entities, so consent forms should "clearly state that this data will be redistributed indefinitely when it becomes a dataset."

ANONYMIZATION AND DATA SECURITY

Experts across disciplines all pointed out the challenge of automatically and fully anonymizing data to make it publicly available without jeopardizing privacy. As one expert cautioned, "Anonymizing audio recordings like transcripts is not easy in any way. It's very likely that the students may have said each other's names or their parents' names." Given that automatic and full anonymization is unlikely, we need to think through carefully if, given the research interests, the intent is to just de-identify or to anonymize completely. For video data, detaching students' faces from their IDs may be sufficient for de-identification, whereas to achieve full anonymization, we will need to blur all faces. Before deciding on an approach for data anonymization, we need to weigh the costs, consider who might bear the costs, and identify which group(s) may be more vulnerable to the accompanying risks. For example, one expert argues that students who rarely speak in class are less likely to be identified compared to "a teacher's voice, who speaks 80% of the time, and [who is] a public figure, and there's only one teacher in the room [compared to many students]."

Similar to data anonymization, there are a lot of options when it comes to data processing and storage approaches. We should keep in mind the end goal and build in flexibility for potential research interests when deciding how the data should be processed and stored. As one interviewee pointed out, "the processing is really variant, depending on the specific experiment […] Most often we will do some […] initial cleaning and the identification as necessary so that the researchers can look at it, but what happens beyond that is really variant, depending on the specific project."

Create synthetic datasets as a possible middle ground for balancing privacy and data availability.

To preserve summary statistics and privacy and avoid an onerous data agreement process, a few experts in the data science and edtech space suggested creating synthetic data that is representative of the underlying data. As one data scientist pointed out, providing a representative synthetic data with key features is helpful for people to try things out and get a sense of what the multimodal data provides, "creating synthetic data that's representative of the underlying data to give people a sense of what's there, and they could try things out. Again, with multimodal data that's more complicated, but, I think, making sure the right people are accessing it that are in line with the intended usage, I think, is fair."

Regardless of the data anonymization and processing approach, to protect privacy, experts from different disciplines agree that PII should be separated from other data pieces as soon as possible in the data processing pipeline. For example, one psychologist shared that for their research project,

they developed an app that "collect[s] the data and transfer[s] audio data directly to a server where the features are directly extracted that we need. And the raw data is deleted." Once data is processed, raw and non-anonymized data should be deleted as early as possible and before the stated timeline on the consent form.

Dissemination Strategy

A key focus of our user case study interviews was to explore potential users' thoughts about the best ways to offer access to the EDSI dataset without compromising privacy, with an emphasis on meeting the needs of R&D in the age of AI. Specifically, we asked: "who should we share this dataset with?" and "what barriers might exist to widespread use of this dataset in the field?" When applicable, we prodded interviewees to share their thoughts on how to ensure broad and equitable access, and how to anticipate the most likely technical and ethical challenges to that access.

While whether and how a dataset like ours can be disseminated depends on specific data usage agreements and consent from participants, interviewees offered several suggestions to guide ours and any data collection effort.

STRATEGIES FOR FIELD ENGAGEMENT AND ADOPTION

Actively and intentionally create a user community.

Experts across fields highlighted that once you've figured out which data components are available in which formats, a designer's focus should be on encouraging broad use of the data in pursuit of generating new knowledge and tools in the field. When asked about specific users we should share the dataset with, individuals readily offered specific intellectual communities, organizations, and individuals that they could see using the data being gathered. While there was a wide array of answers here, almost all respondents agreed that regardless of discipline or background, there is a need to intentionally create awareness of the dataset.

The underlying logic that was shared repeatedly is that active steps are needed to increase awareness, utilization, and continued engagement with the data. One AI researcher put it best, "I think to build [a user] community [...] there has to be some incentives [to engage with the data]", with almost all experts suggesting that an intentional approach focusing on reaching out to key stakeholders facilitates the creation of user communities and could help demonstrate the utility of the dataset, compounding usage. That same AI researcher said the adage "if you build it, they will come" does not often work in practice, and highlighted that "there's got to be some structure to get people to actually work on it, and to create a field awareness that it's there and it can be used for good stuff." Respondents emphasized that these efforts need not be complicated; they can be as simple as engaging key people within communities of interest to "translate it in a way that's perceptible and understandable to that community" by explicitly sending this project and idea to research groups to work with the dataset.

The efforts to build user communities should begin early and happen often.
As the data are being collected, practically all interviewees stressed the importance of sharing the data early and broadly across user groups. These efforts will not only raise awareness, but also potentially increase the usability of the data by for example troubleshooting syncing or other processing issues and/or identifying adjustments to the data collection or measures to allow for additional major use cases. One AI researcher summarized this idea saying, "only by engaging with the community [...] you probably will get all of this information better." Importantly, experts suggested the idea of documenting and sharing these initial users' analytic plans and annotation strategies/efforts to provide examples of use cases and potentially non-examples of the dataset across various fields.

ETHICAL CONSIDERATIONS FOR SHARING

In addition to offering dissemination recommendations, experts highlighted key ethical considerations that should be considered when sharing sensitive data of our nature. Most of the interviewees focused our attention on the inherent tradeoff between user access and data security that any responsible designer of a dataset should attempt to balance. Potentially complicating this balancing act is the learning that different groups have different tolerances for various types of security features: One AI researcher noted "[if] you want to design something that's accessible [for advanced methods, the data] needs to have some degree of being able to download it", highlighting how downloading the data and having it on personal computers is seen as a benefit for access, while social science and education researchers emphasized a malleable secure environment that facilitates various types of analysis for identifiable data. Prior to providing any individual access to the data, experts noted the need to document with detail who has access to which types of data. As one social scientist phrased it, "Who can access? When? How? For how long?" Most experts noted that knowing who has access is the first step in ensuring that the data are being accessed by actors with good intentions – further protecting the privacy of the students and teachers present in the data.

Some experts offered the suggestion of developing a steering committee that represents diverse needs to ensure accessibility for a wider variety of audiences. Others encouraged us to host workshops on how to work in whatever workspace we provide, especially for higher security infrastructure that can be cumbersome to navigate. One experienced AI researcher reflected:

> *One of the things that I've seen in the past with the MET study and with others is that they offered workshops for regular engagement and training researchers how to access and use these data […] Offering opportunities to learn about the data set and to engage with one of the original developers or researchers that collected the data, I think that would be hugely beneficial both […] for inviting people into it, but then also spreading the word, and then also not just spreading the word, but providing directly accessible pathways to using the data.*

FLEXIBLE ACCESS PATHWAYS

There is no perfect solution that will accommodate all types of analysis.

Given the multiple sources of data, wide array of potential use cases, and the variation in technical skills and computing needs, there is not a uniform secure data access plan that covers the entire corpus of data for the EDSI dataset or similar endeavors. Instead, designers must calculate their tolerance for privacy risk against usability for differing levels of sensitive data within the boundary of their legal obligations– taking steps to work with specific communities to ensure the right data are accessible in the format that will generate the maximum amount of new knowledge/tools, while protecting student and teacher privacy

Recognizing this, experts encouraged a tiered system based on the level of data sensitivity. In other words, they suggested that we have different versions of the data available via different methods, with the levels of access depending on the level of PII present. This suggestion occurs in practice, with one social scientist stating "I know systems where you have, for instance, 3 levels [with increasingly more secure ways of accessing]." Specific recommendations as it pertains to our dataset are as follows:

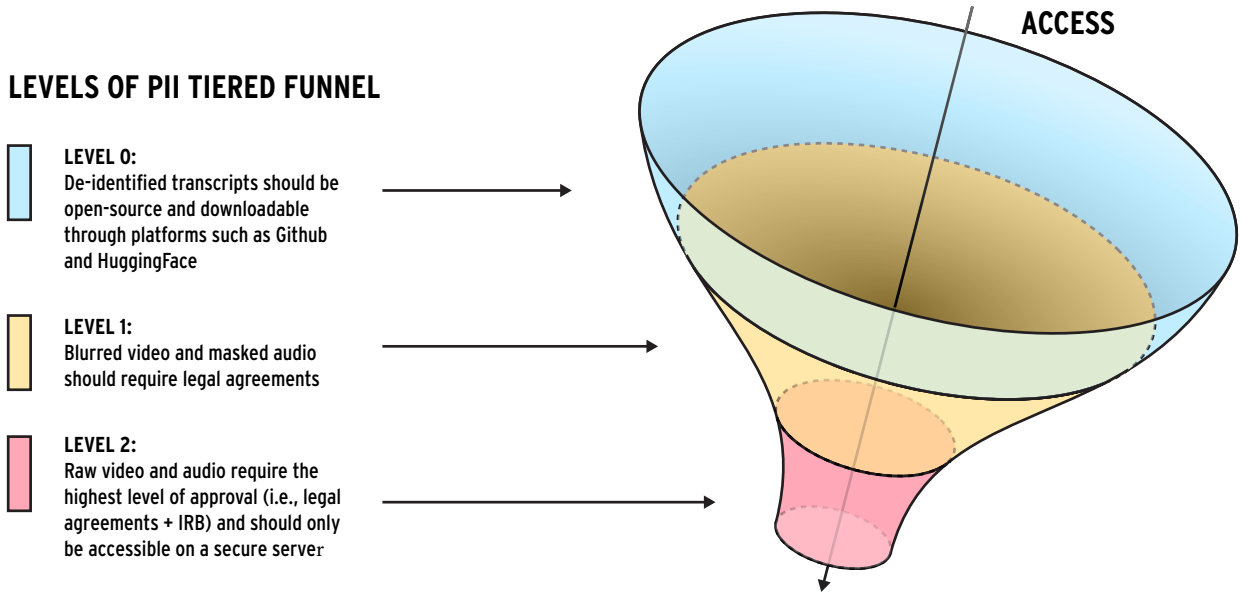

Figure 4: Access Funnel

A well designed and malleable computing infrastructure facilitates ease of use.

There is a clear tradeoff between the security level of data access and the need for providing computing. One interviewee highlighted that the "key for this kind of data, is you've got to have the data de-identified to a degree that's trusted, or you have to have an infrastructure more that allows execute, access, downloading." In fact, one of the most prominent things we heard in interviews is that in addition to ensuring that the requested types of data are available, the environment in which they access them must be adaptable for many purposes. A specific example is ensuring that the data environment has the appropriate software and tools that individuals will need to complete their analysis, which poses a particular challenge given that most AI models

require internet access, how fast AI models evolve, and how data are handled by AI companies. Alternatively, some experts suggested hosting the data on a platform like DataShop or creating a package such that users can see aggregate outputs only but are still able to complete their analysis.

Secure data environments can be challenging for certain users.

In order to ensure broad equitable access to this dataset, there are particular challenges that should be anticipated and addressed in the secure data access design. Many interviews noted that more secure data environments such as enclaves "are really great for experienced power users, but not great [for more novice users].. " Researchers who are at the forefront of AI research also noted these tradeoffs between access and flexibility and security, sharing that:

> *Everyone works [differently]. They work in different languages, they work in different modes and different workflows. And so you severely restrict the diversity and the creativity with which people can approach a problem if they don't have access to different resources and platforms that they're kind of used to.*

Proactive strategies to ensure various types of users can work with the data in a secure environment included i) providing access to software and coding resources with the goal of "taking away the technological barriers [...] and making sure that researchers are coming into a familiar coding environment for them"; ii) offering adequate storage and strong computing power for those who require it; and iii) sharing measures others have created along with the features we extract to facilitate similar types of analysis because, as one social science researcher put it "[if] somebody else who's also interested in [studying] student participation doesn't have to do the same analysis."

Bureaucratic barriers in higher education institutions may compromise wider use.

One data scientist noted that the lack of access to an IRB approval can serve as a barrier for some folks outside academia. Some experts suggested that we publish secured aspects of the dataset in places that are easily accessible for those in industry. Specific examples mentioned included posting code on GitHub or models on HuggingFace, or hosting a Kaggle competition for particular research questions to boost access by people in the machine learning AI community.

Particularly for junior researchers, graduate students , or other folks without connections, incentives or funding support can help . One senior AI researcher leveled the playing field by bootstrapping the data with their own team of grad students and then got funding to open it up to the community of researchers and create small grants for them. Competitions and grand challenge models can also incite interest and provide needed funding for these groups.

A Final Word

The insights from our expert interviews have laid a clear path forward for the EDSI project, affirming that high-quality, purpose-built data are the cornerstone for fulfilling the promise of AI in education. Intentionally designed, multimodal datasets that capture the nuanced dynamics of the classroom, prioritize crisp audio, and account for contextual factors are essential for meaningful innovation. Such datasets can also help advance key research initiatives in both teaching and learning and technical breakthroughs in AI in education. We have also learned a great deal about privacy and confidentiality concerns and solutions, appreciating the changing nature of this highly sensitive topic due to constant technology development. As we continue to build this new benchmark dataset, we are guided by a commitment to transparency, ethical data handling, and broad accessibility. We believe the EDSI dataset, informed by these principles, will be more than just a collection of information-it will be a shared resource that accelerates R&D, democratizes access to powerful tools, and ultimately helps the entire field collectively build a future where AI genuinely augments human expertise to support effective teaching and student success.

EDSI

CENTER FOR EDUCATIONAL
DATA SCIENCE & INNOVATION

UNIVERSITY OF MARYLAND

Appendix

EXPERT INTERVIEW PROTOCOL

The protocol below was used for interviews with researchers specializing in education and social science research pertinent to AI development. All interviews included the same anchor questions, but other interview groups (i.e. solution developers, data scientists, and psychologists) substituted some of the probe questions bulleted beneath each anchor question with questions appropriate to their expertise. Because interviews were semi-structured, the interviewer asked additional probing questions as necessary and at times asked questions out of order as the conversation demanded. All interviews lasted approximately 60 minutes.

[INTERVIEWER]

We are going to start by asking some questions to understand your previous experience working with multimodal datasets, and your perceptions of AI in education. Our goal is to spend about 10 minutes in this section.

Prior Experience with Multimodal Data

Q1a: Have you worked with multimodal datasets that utilize audio and video recordings before?

- If yes, What worked well? What challenges did you encounter?
- Have you encountered challenges in ensuring data quality and consistency in multimodal datasets?
- How do you typically preprocess multimodal datasets before analysis?
- What limitations have you faced when working with similar datasets?

Perceptions of AI in Education

Q1b: What are some of the biggest opportunities AI presents in educational research and practice?

- What are some promising AI applications in educational research?

Q1c: What concerns or risks do you see with AI in education?

- Are there key limitations or risks you see in AI-driven educational insights?

Now, we are going to move on to questions that explore your reactions to the high-quality multimodal dataset that EDSI is developing, asking questions about the 1) features and characteristics, 2) implications for the field and 3) any ethical / privacy concerns. Our goal is to spend about 30 minutes in this section.

Features and Characteristics

Q2a: What features stand out to you in this dataset?

- How does this dataset compare to others you've worked with?

Q2b: What features or components would you like to see added?

- We are still thinking about administering a baseline teacher survey. What questions would you ask teachers?
- What specific data structures, formats, or annotations would make this dataset more usable for research?

Q2c: What aspects of quality do you see as particularly important in datasets of this nature?

- What additional metadata or documentation would be helpful to make the dataset more useful?
- Are there particular variables or metadata that would enhance its research value?

Implications for the Field

Q2d: How would you use this dataset?

- Are there particular research questions or applications that excite you?

Q2e: How might other stakeholders in your field use this dataset?

- …

Q2f: What are the best uses of this dataset (e.g., training models, benchmarking, developing new tools)?

- How could this dataset be leveraged to improve existing AI models in education?

Q2g: Do you think this dataset can benefit the field of AI and education?

- If yes, In what ways?

Ethical/Privacy Concerns:

Q2h: Are there any limitations or potential biases in this dataset that concern you?

- …

Q2i: Do you have any concerns about privacy, security, or ethical considerations regarding the collection and use of this dataset?

- What are best practices for ensuring fairness and transparency in AI-based educational research?
- What ethical considerations should we account for in dataset sharing and access?

Q2j: Who should we share this dataset with?

- What ethical considerations should we keep in mind when sharing the dataset?

Thank you. Now we are going to move onto our final questions - addressing other considerations related to access and our conversation. In total, we will spend about 7.5 minutes in this section.

Other Considerations

Q3a: What barriers might exist to widespread use of this dataset in the field?

- How should we ensure broad, equitable access to this dataset?
- Are there technical challenges (e.g., interoperability, data size, missing values) that could limit usage?
- How could we improve dataset documentation to maximize accessibility?

Q3b: Are there any questions we aren't asking? Or points you haven't yet made we need to hear?

- …

CLOSING

That concludes our interview. You will receive your $150 Visa gift card within one month of today's date. Thank you for agreeing to participate in our study!

We look forward to sharing our findings with you, and as one final request, we would like to know if there are any {stakeholders} that we should speak with (based on our conversation)?